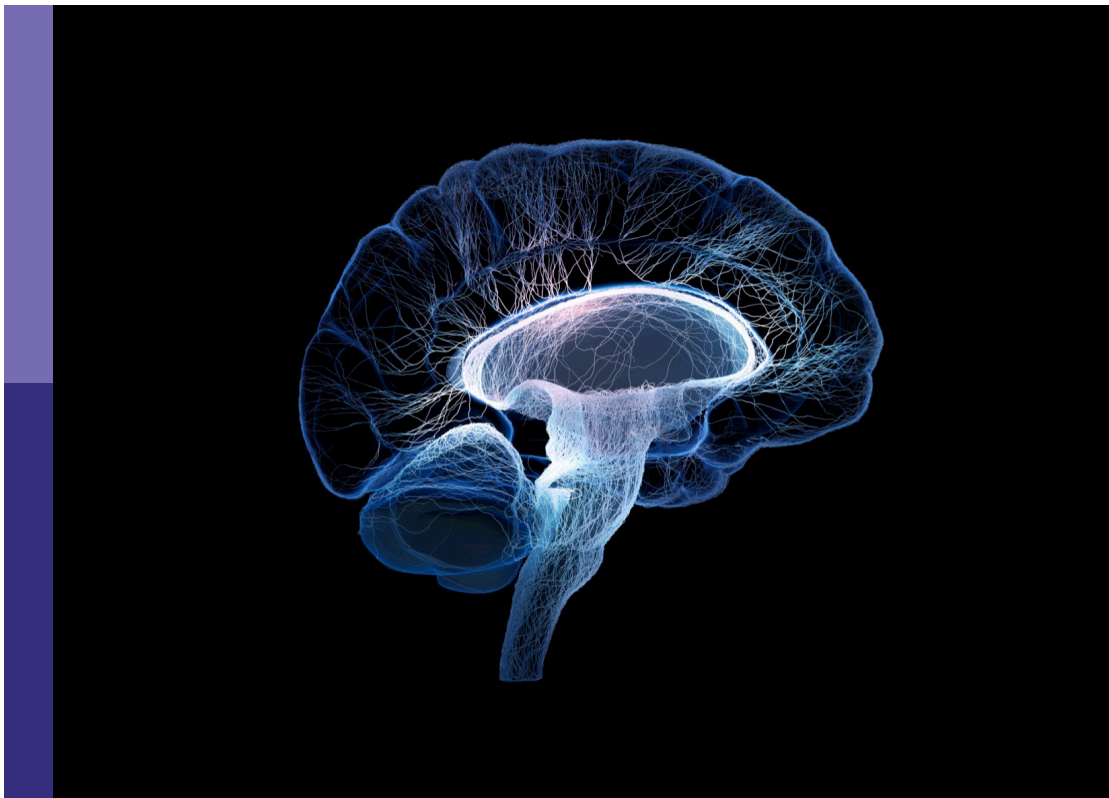
frontiers | Research Topics

Advanced machine learning approaches for brain mapping

Edited by
Dajiang Zhu, Shu Zhang, Xi Jiang and Dingwen Zhang

Published in
Frontiers in Neuroscience
Frontiers in Neuroimaging



ISSN 1664-8714
ISBN 978-2-8325-4757-1
DOI 10.3389/978-2-8325-4757-1

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

Advanced machine learning approaches for brain mapping

Topic editors

Dajiang Zhu — University of Texas at Arlington, United States
Shu Zhang — Northwestern Polytechnical University, China
Xi Jiang — University of Electronic Science and Technology of China, China
Dingwen Zhang — Northwestern Polytechnic University, United States

Citation

Zhu, D., Zhang, S., Jiang, X., Zhang, D., eds. (2024). *Advanced machine learning approaches for brain mapping*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4757-1

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 22 November 2022
DOI 10.3389/fnins.2022.1046268

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Shijie Zhao,
Northwestern Polytechnical University, China
Yuqi Fang,
University of North Carolina at Chapel Hill, United States

*CORRESPONDENCE
Xiaoke Hao
haoxiaoke@hebut.edu.cn
Jing Qin
harry.qin@polyu.edu.hk

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

RECEIVED 16 September 2022
ACCEPTED 02 November 2022
PUBLISHED 22 November 2022

CITATION
Hao X, An Q, Li J, Min H, Guo Y, Yu M and Qin J (2022) Exploring high-order correlations with deep-broad learning for autism spectrum disorder diagnosis.
Front. Neurosci. 16:1046268.
doi: 10.3389/fnins.2022.1046268

Exploring high-order correlations with deep-broad learning for autism spectrum disorder diagnosis

Xiaoke Hao[1]*, Qijin An[1], Jiayang Li[1], Hongjie Min[1], Yingchun Guo[1], Ming Yu[1] and Jing Qin[2]*

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China, [2]School of Nursing, Centre for Smart Health, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

Recently, a lot of research has been conducted on diagnosing neurological disorders, such as autism spectrum disorder (ASD). Functional magnetic resonance imaging (fMRI) is the commonly used technique to assist in the diagnosis of ASD. In the past years, some conventional methods have been proposed to extract the low-order functional connectivity network features for ASD diagnosis, which ignore the complexity and global features of the brain network. Most deep learning-based methods generally have a large number of parameters that need to be adjusted during the learning process. To overcome the limitations mentioned above, we propose a novel deep-broad learning method for learning the higher-order brain functional connectivity network features to assist in ASD diagnosis. Specifically, we first construct the high-order functional connectivity network that describes global correlations of the brain regions based on hypergraph, and then we use the deep-broad learning method to extract the high-dimensional feature representations for brain networks sequentially. The evaluation of the proposed method is conducted on Autism Brain Imaging Data Exchange (ABIDE) dataset. The results show that our proposed method can achieve 71.8% accuracy on the multi-center dataset and 70.6% average accuracy on 17 single-center datasets, which are the best results compared with the state-of-the-art methods. Experimental results demonstrate that our method can describe the global features of the brain regions and get rich discriminative information for the classification task.

KEYWORDS

autism spectrum disorder, high-order functional brain network, broad learning system, classification, feature selection

Introduction

Autism spectrum disorder (ASD) is a neurologically heterogeneous disorder that is difficult to diagnose. The main characteristics of ASD patients are social interaction disorders and neurodevelopmental disorders of stereotyped behavior. The life expectancy of ASD patients is much lower than that of normal controls (NC) (Perkins and Berkman, 2012). The current psychiatric diagnosis for ASD refers only to symptomatic behavioral observations (DSM-5/ICD-10), which may be misdiagnosed (Nickel and Huang-Storms, 2017). However, the cause and pathogenesis of ASD are unclear. There is an urgent need to identify biomarkers associated with brain imaging data to assist medical diagnosis.

Recently, various non-invasive brain imaging techniques such as magnetic resonance imaging (MRI), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) are widely used in the study of neurodegenerative diseases such as ASD. In particular, several studies in recent years have shown that using the blood oxygen level-dependent (BOLD) signal as a neurophysiological indicator can effectively identify potential biomarkers in ASD patients (Dekhil et al., 2018). Many studies have been conducted based on low-order brain functional connectivity obtained from fMRI, which reflects the correlation relationship between signals from paired brain regions (Liang et al., 2012; Li et al., 2014). Due to the spontaneous aberrations generated in the functional connectivity status of brain disease patients (Hahamy et al., 2015), there is a significant variability compared to NC.

Many studies explore the low-order brain functional connectivity to diagnose ASD. Stacked multiple sparse autoencoders (SSAE) is applied to learn the discriminative feature representation of low-order brain functional connectivity and subsequently diagnose ASD (Kong et al., 2019). Dekhil et al. (2018) construct an ASD diagnostic system consisting of sparse autoencoders and spatially activated regions, which similarly learn low-order brain functional connectivity features. Wang et al. (2020) propose a multi-site domain adaptation method based on low-order brain network for ASD diagnosis. Wang et al. (2022) propose a multi-site clustering and nested feature extraction method for fMRI-based ASD detection. However, current methods only reflect correlations between pairs of brain regions. The connections between brain regions are complex, and studies that only reflect pairwise relationships between brain regions are still limited. In contrast to the traditional approaches to characterize lower-order brain functional connectivity (Yu et al., 2017; Li et al., 2019), high-order feature representation of brain connectivity can characterize complex patterns of interactions between multiple brain regions and correlations across brain regions. Feng et al. (2020) regard the second-order functional connectivity network as a higher-order brain network, in which brain connectivity patterns are only obtained by repeatedly computing first-order correlations between pairs of brain regions, and some features may be lost in the process of repeated computing. Gao et al. (2020) exploit clustering of functional connectivity time series to reveal high-order relationships among multiple regions of interest (ROIs), but global brain functional connectivity features have not been considered.

To overcome the above-mentioned drawbacks and to form high-order feature information that can characterize the global structure of the brain, we introduce the hypergraph structure to inscribe the high-order brain functional connectivity. Hypergraph (Ktena et al., 2018) is a novel tool for inscribing high-order structures, and the features of the hypergraph structure are distinguished from the traditional graph structure features. Unlike normal graphs, hypergraphs are composed of nodes and hyperedges. One hyperedge can connect two or more nodes. Hypergraph learning is flexible and powerful in modeling complex data dependencies such as brain networks. It has received more attention that using hypergraph to describe the brain connection pattern can more accurately describe the complex high-order connection relationship of the brain network.

Due to the complex features of brain networks, several recent studies use deep learning-based approaches to diagnose patients with autism spectrum disorders (Ktena et al., 2018; Gao et al., 2020; Yao et al., 2021). For example, Eslami et al. (2019) propose a self-encoder-based model for ASD classification. Heinsfeld et al. (2017) use two stacked denoising autoencoders to identify ASD patients from fMRI data. Xing et al. (2019) propose a novel convolutional neural network with elemental filters for the diagnosis of ASD. Huang et al. (2021) use Long Short-Term Memory Networks (LSTM) for the classification of ASD patients. Guo et al. (2017) and Khodatars et al. (2021) propose a deep neural network model for the study and diagnosis of ASD patients. Zhang et al. (2022) propose a feature selection method based on variational autoencoder pre-training using a multilayer perceptron for ASD classification. Jiang et al. (2020) propose a hierarchical GCN framework to learn brain network graph feature embeddings while considering both network topology information and subject associations. However, all these deep learning-based methods and graph neural network-based methods are based on low-order brain functional connectivity networks for subsequent feature extraction and classification, which have non-negligible drawbacks. Firstly, there are limitations in using low-order brain network features to represent brain connectivity patterns. Secondly, the models based on deep learning will become more complex as the number of model layers increases, the training process is time-consuming and the deep network features are not scalable. The number of parameters to be learned is huge, and it often faces the problem of insufficient single-center data, resulting in overfitting. It is not until the emergence of the Broad Learning System (BLS) (Chen and Liu, 2018) that

traditional artificial intelligence methods are revolutionized. It represents a step towards building more effective machine learning methods that can further extend models based on deep learning methods and improve the learning efficiency of the models (Chen and Liu, 2018; Gong et al., 2022). Recently, some studies have introduced BLS and its variant algorithms into medical image analysis (Han et al., 2020), providing an effective tool for diagnosing AD in MRI images. However, there are no studies based on BLS to diagnose ASD in fMRI. Benefiting from the superiority of BLS, we use it for further feature selection and classification.

Compared with traditional deep learning-based diagnostic models, our proposed deep-broad learning method can learn complex and high-dimensional brain connectivity network features more accurately. We use the functional connectivity and hypergraph structure of fMRI to characterize the high-order connectivity characteristics of the brain. The feature learning process is further extended by using the structure fused by the autoencoder and the BLS, and an efficient and accurate brain network learning structure is obtained. The main contributions of this paper are summarized as follows:

Firstly, we construct a high-order brain functional connectivity network of the functional connectivity structure of fMRI based on hypergraph structure, which improves the ability of traditional brain functional connectivity networks to express brain structure.

Secondly, we propose a novel combinatorial deep-broad learning method to extract high-dimensional discriminative features of high-order brain functional connectivity networks.

Compared with other ASD classification models, our model not only takes into account the global functional connectivity features of the brain but also provides a feature-learning classification module with fewer parameters using BLS.

The purpose of this paper is to propose an effective model for portraying the global functional connectivity structure of the brain. The BLS further enhances the feature learning capability and computational speed of deep learning models for ASD diagnosis. The rest of the paper is structured as follows. In section 3, we introduce the dataset materials and the details of our proposed method. In section 4, we perform an experimental evaluation and experimental analysis of the proposed method. In section 5, we summarize the work presented in this paper.

Materials

In this study, we use 505 patients with ASD and 530 healthy controls from 17 sites in the ABIDE- I [ABIDE (http://fcon_1000.projects.nitrc.org/indi/abide/)] dataset for our experiments. Our study uses data pre-processed by the C-PAC pipeline (Agastinose Ronicko et al., 2020) with the following pre-processing processes: motion correction, slice timing correction, removal of interfering signals, low-frequency drift and voxel intensity normalization. ABIDE provides a variety of ROI segmentation options. In this study, we use 200 uniform ROIs generated by the spatially constrained spectral clustering algorithm (Craddock et al., 2012).

Method

We propose a deep-broad learning method to explore high-order brain functional connectivity network features for ASD classification. The specific structure of the model is shown in **Figure 1**. The model consists of four parts. (1) Firstly, the low-order brain functional connectivity network is constructed by calculating the time-series Pearson correlation matrix of the fMRI data. It is used to portray the low-order local features of the brain shown in **Figure 1** as Lo-FCN. (2) We introduce the hypergraph structure to construct high-order brain functional connectivity network to inscribe the global features of the brain shown in **Figure 1** as Ho-FCN. (3) Initial feature learning of high-dimensional high-order brain functional connectivity network is performed using an autoencoder. (4) Finally, the initial features learned by autoencoder are fed into the BLS for further learning and classification. The details of each step will be given in the following sections.

Construction of high-order brain functional connectivity network

We obtain the high-order brain functional connectivity network based on time series of fMRI and coactivation level signals based on hypergraph to effectively characterize the global brain connectivity pattern. Specifically, we first calculate the correlations between pairs of brain regions using Pearson correlation coefficients, which are widely used to calculate the functional connectivity of fMRI (Liang et al., 2012; Baggio et al., 2014; Zhang et al., 2017), as shown in Equation (1). u and v represent the time series of two ROIs, the length of each series is T, $\overline{u}$ and $\overline{v}$ represent the average values of the time series u and v, respectively. Calculating the pairwise correlation of all time series will get the paired brain regions correlation matrix $Corr_{M \times M}$, where M is the number of ROIs, so as to obtain the low-order brain functional connectivity network.

$$\rho_{\mathbf{uv}} = \frac{\sum_{\mathbf{t}=1}^{\mathbf{T}} (\mathbf{u_t}-\overline{\mathbf{u}})(\mathbf{v_t}-\overline{\mathbf{v}})}{\sqrt{\sum_{\mathbf{t}=1}^{\mathbf{T}} (\mathbf{u_t}-\overline{\mathbf{u}})^2}\sqrt{\sum_{\mathbf{t}=1}^{\mathbf{T}} (\mathbf{v_t}-\overline{\mathbf{v}})^2}} \quad (1)$$

Each element in the low-order brain functional connectivity matrix depicts the correlation between local pairwise ROIs. When the paired ROIs are highly correlated, the element value approaches 1, and when the paired brain interval is inversely correlated, the element value is close to -1. Since pairwise relationships between brain regions only characterize

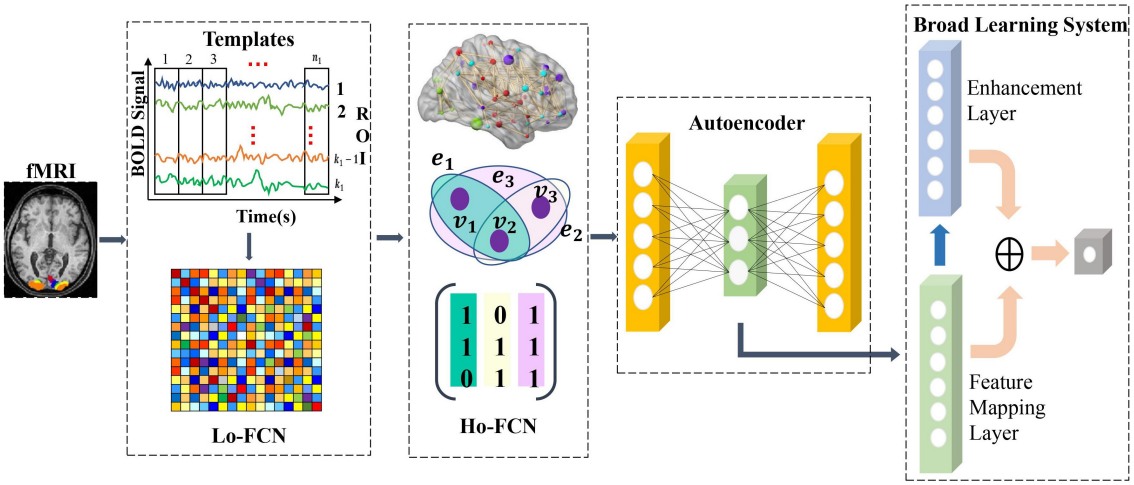


FIGURE 1
The illustration of our proposed method.

local features of the brain, we use the hypergraph to represent correlations in the interaction of multiple brain regions rather than pairwise correlations, resulting in a hypergraph-based high-order brain network. Specifically, we first recall the basics of hypergraph (Schölkopf et al., 2007), where we denote a hypergraph as $G = \{V, E, W\}$, the set of hypergraph vertex of the hypergraph $V = \{v_1, v_2, ..., v_n\}$ represents the n brain regions, hyperedge $E = \{e_1, e_2, ..., e_m\}$ represents the correlation between brain regions, each hyperedge is assigned a weight $w(e_i), 1 \leq i \leq M$. The weight vector of a hyperedge is expressed as $W = \{w_{e_1}, w_{e_1}, ..., w_{e_m}\}$. The hypergraph structure can be represented simply as an association matrix $H \in \{0, 1\}^{|V| \times |E|}$, each element in H indicates whether the vertex v is in the hyperedge e,denoted as $H(v, e) = \begin{cases} 1 \; if \; v \in e \\ 0 \; if \; v \notin e \end{cases}$. The elements in the association matrix represent the probability value of the importance of the node to the hyperedge. Based on the constructed association matrix, the degree of the hypergraph node and the degree of the hyperedge can be obtained, which are expressed as: $d(v_i) = \sum_{e_j \in \varepsilon} w_{e_j} H_{ij}$ for $1 \leq i \leq N, \delta(e_j) = \sum_{v_i \in v} H_{ij}$ for $1 \leq j \leq M$. The degree matrix of the hypergraph vertex and the hyperedge are described as $Degree_e \in R^{|E| \times |E|}$ and $Degree_v \in R^{|V| \times |V|}$. $D_e \in R^{|E| \times |E|}$and $D_v \in R^{|V| \times |V|}$ are the diagonal matrices containing the hypergraph vertex and the hyperedge. In graph theory, the graph Laplacian matrix plays an important role in graph learning. Based on the graph Laplacian matrix, by calculating its eigenvalues and eigenvectors, According to previous research (Schölkopf et al., 2007), we can perform a spectral analysis of the graph. For simple graphs, the graph Laplacian is defined as $\triangle = D - A$, D is the diagonal matrix of vertex degrees, and A is the adjacency matrix, while for hypergraphs, the graph Laplacian is defined as $\triangle = D_v - HWD_e^{-1}H^T$. The normalized Laplace matrix is $\triangle = I - D_v^{-\frac{1}{2}} HWD_e^{-1}H^T D_v^{-\frac{1}{2}}$. We summarize the symbols and definitions in Table 1.

In order to obtain a hypergraph-based high-order brain functional connectivity network, we construct the hypergraph using the method proposed by the previous work (Schölkopf et al., 2007). We treat each hypergraph node as a brain region and use each brain region as a central region to calculate the Euclidean distance between the selected central region and other brain regions. Specifically, we first take each vertex (ROI) as a center node and calculate the Euclidean distance between the center and other vertices. Then we construct a hypergraph by connecting the center and its K nearest vertex. We regard the k brain regions closest to the central node Euclidean space $d_{ij} = |v_i - v_j|_2^2$ as the nearest neighbors of central brain region. We refer to correlations between nodes as a hyperedge, d_{ij} represents the Euclidean distance between brain region v_i and

TABLE 1 Symbols and definitions of hypergraph.

Notation	Definition
$G = \{V, E, W\}$	G represents the hypergraph, V, E, W represent the set of vertices, the set of hyperedges, and diagonal matrix of hyperedge weights. respectively.
$d(v_i)$	The degree of vertex v_i.
$w(e_i)$	The weight of hyperedge e_i.
$\delta(e_j)$	The degree of hyperedge e_j.
H	The $\|V\| \times \|E\|$ incidence matrix of hypergraph structure. $H(v, e)$ indicate the connection strength between the vertex v and the hyperedge e.
D_v	The diagonal matrix of vertex degrees.
D_e	The diagonal matrix of hyperedge degrees.
$\triangle$	The Laplacian matrix of hypergraph.

v_j. Based on the above mentioned process, we construct the hypergraph based high-order brain connectivity network to represent the global features of brain.

The novel feature extraction and classification method based on deep-broad learning

Due to the high dimensionality of the constructed high-order brain functional connectivity network features, we utilize a non-linear dimensionality reduction approach to reduce the feature dimension. Specifically, we use an autoencoder to reconstruct the features of the constructed hypergraph-based high-order brain feature representation. We obtain low-dimensional discriminative features of the high-order brain network by minimizing the reconstruction error between the input network features and the output features through self-supervised learning. The autoencoder consists of an encoder and a decoder.

$$_{enc} = \emptyset_{enc}(x) = \tau\left(W_{enc}x + b_{enc}\right) \quad (2)$$

We use the original high-order brain feature representation x as input to the autoencoder to obtain discriminative lower-dimensional feature h_{enc} via the encoder, denoted as Equation (2) where τ is the hyperbolic tangent activation function (*tanh*), and W_{enc} and b_{enc} represent the weight matrix and bias of encoder. Once we have obtained the low-dimensional feature representation h_{enc} of the high-order brain function connectivity network, we use the decoder to reconstruct the original input data x, expressed as Equation (3). The low-dimensional feature representation h_{enc} is input into the decoder, where W_{dec} and b_{dec} represent the weight matrix and bias of the decoder, respectively. We use Mean Squared Error (MSE) as the reconstruction loss, which represents the discrepancy between the reconstructed brain function connectivity network features x' and the original features x. After completing the training of the autoencoder, we obtain the low-dimensional feature representation of the new high-order brain functional connectivity network as the effective feature. And we use it as the valid discriminative high-order brain function connectivity feature input broad learning system for further learning.

$$x' = \emptyset_{dec}\left(h_{enc}\right) = W_{dec}h_{enc} + b_{dec} \quad (3)$$

Analyzing higher-order brain functional connectivity using existing machine learning methods is challenging due to the high-dimensional, large-scale, and complex interdependencies between brain regions. Moreover, a large number of iterative processes during traditional model training requires huge amounts of time and computational resources. The Broad Learning System (BLS) (Chen and Liu, 2017, 2018; Chen et al., 2019) has recently become one of the most popular networks due to its excellent performance in machine learning tasks (Gong et al., 2022). BLS can map samples to a more suitable space to handle the large volume of high order brain functional network features and is suitable for processing time-varying data. BLS first map the inputs to construct a set of mapped features. A group of mapping nodes defined in our work is a mapped feature in original BLS. Given that the feature extraction at this step uses randomly generated weights, calculating multiple mapping features can enhance the stability of the extracted feature information and simplify the operations. **Figure 2** illustrates the basic structure of the BLS, which consists of a three-layer network defined as the feature mapping layer, the enhancement layer and the output layer, where $X \in R^{N \times m}$ denotes the discriminative high-order brain function connectivity matrix learned by autoencoder, which is taken as the input to the BLS. N is the number of samples, and m is the feature dimension of each sample, $Y \in R^{N \times c}(c < m)$ is the output layer of BLS, c is the feature dimension after the feature extraction by BLS of each sample, and W_{BLS} is the weight of the feature mapping layer and the enhancement layer to the output layer. Specifically, we first input the high-order brain function connectivity matrix X to the feature mapping layer to generate the i-th group of mapping nodes Z_i, denoted as Equation (4):

$$Z_i = \emptyset(XW_{ei} + \beta_{ei}), i = 1, \ldots, n \quad (4)$$

W_{ei} and β_{ei} are the weight and bias from X to the feature mapping layer. Similarly, the *m-th* group of enhancement layer nodes H_m is generated by taking the mapping node as the input of the enhancement layer, which is expressed as Equation (5):

$$H_m\delta \equiv \left(Z^n W_{h_m} + \beta_{h_m}\right) \quad (5)$$

W_{h_m} and β_{h_m} are the weights and biases from the feature mapping layer to the enhancement layer. It should be noted that $\emptyset$ and δ are nonlinear functions, such as *tanh* and *tansig*,

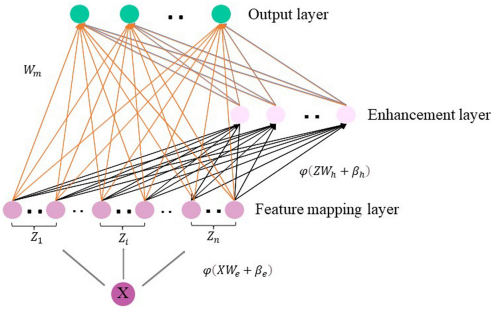


FIGURE 2
The structure of the broad learning system.

TABLE 2 Classification performance of different methods on multi-centers dataset.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Running time(s)
SVM	60.3 ± 3.6	35.3 ± 8.2	84.4 ± 6.6	64.3 ± 5.9	186
Random forest	63.7 ± 7.2	54.9 ± 3.9	71.3 ± 9.2	68.9 ± 3.8	67
DNN	60.7 ± 5.4	56.4 ± 7.3	64.8 ± 3.8	70.1 ± 5.6	108030
Autoencoder	65.4 ± 6.6	69.3 ± 4.2	61.9 ± 5.2	60.8 ± 3.7	21600
Ours	71.8 ± 4.2	70.8 ± 3.8	65.9 ± 4.2	65.9 ± 4.8	1200

we compose all mapping nodes as $Z = [Z_1, Z_2, ..., Z_n]$, and enhance the nodes as $H = [H_1, H_2, ..., H_m]$.

The BLS model can be expressed as Equation (6):

$$Y = \left[Z_1, Z_2, ..., Z_n | \delta\left(Z^n W_{h_1} + \beta_{h_1}\right), ..., \delta\left(Z^n W_{h_m} + \beta_{h_m}\right)\right]$$
$$W_{BLS} = [Z_1, Z_2, ..., Z_n | H_1, H_2, ..., H_m]\, W_{BLS}$$
$$= [Z|H]\, W_{BLS} \quad (6)$$

To summarize, we further learn the features extracted by autoencoder via BLS and finally get the ASD classification result.

Experiments

In this section, we conduct two-stage experiments of our proposed method. In the first stage, we conduct experiments on 1,035 samples from 17 multi-centers and each single-center to demonstrate the effectiveness of our proposed method. In the second stage, we compare with the state-of-the-art methods using another atlas that divides the brain into 264 ROIs. The robustness and scalability of our proposed method are further verified by experiments on multi-atlas data.

TABLE 3 Classification performance of our proposed method on single-center dataset.

Sites	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
NYU	71.4 ± 5.7	75.0 ± 1.5	68.4 ± 9.4	66.7 ± 9.8
OHSU	80.8 ± 3.8	76.9 ± 7.7	84.6 ± 7.7	83.3 ± 8.3
KKI	62.5 ± 8.3	63.6 ± 9.1	61.5 ± 7.7	58.3 ± 8.4
YALE	71.4 ± 3.6	74.0 ± 2.9	68.9 ± 4.4	68.9 ± 2.5
USM	81.6 ± 2.9	84.8 ± 4.1	76.0 ± 0.9	86.6 ± 0.3
Olin	76.4 ± 3.0	75.0 ± 3.9	78.6 ± 1.4	83.3 ± 0.0
Pitt	73.2 ± 1.8	75.0 ± 2.6	71.4 ± 6.4	72.4 ± 5.4
Leuven	74.6 ± 3.2	68.8 ± 5.4	80.6 ± 0.7	78.6 ± 0.7
UCLA	79.5 ± 3.1	77.8 ± 3.7	75.0 ± 2.3	79.2 ± 2.3
Caltech	67.6 ± 2.7	71.4 ± 0.0	62.5 ± 6.3	71.4 ± 3.6
CMU	66.7 ± 3.6	64.3 ± 7.1	69.2 ± 0.0	69.2 ± 2.2
MaxMun	65.4 ± 3.8	69.2 ± 0.3	61.5 ± 6.7	64.3 ± 4.9
SBL	66.7 ± 3.4	71.4 ± 1.2	62.7 ± 6.2	62.5 ± 4.9
SDSU	61.6 ± 2.7	60.1 ± 5.7	67.9 ± 3.8	65.6 ± 3.2
Stanford	67.7 ± 4.2	51.6 ± 5.7	76.9 ± 5.3	64.2 ± 3.7
UM	63.8 ± 5.6	70.5 ± 4.3	53.2 ± 5.7	75.6 ± 4.9
Trinity	68.8 ± 5.5	72.9 ± 4.6	60.5 ± 5.7	73.3 ± 3.7
AVERAGEE	70.6 ± 4.0	71.7 ± 4.1	69.4 ± 4.8	71.7 ± 4.0

In our experiments, we use 10-fold cross-validation to evaluate the classification accuracy of the prediction model. This means that we first randomly divide the dataset into 10 disjoint subsets of data, and then select a single subset as the test set, with the remaining 9 subsets used as the training set. In particular, for multi-center experiments, we mixed all samples from 17 centers, and then divided the dataset into 10 disjoint subsets. We selected 1 of the K_e parts as the test set and the remaining K_e-1 parts as the training set, finally took the average of the K_e verification results as the verification error of this model. The process is repeated 10 times to reduce the effect of sampling bias on the experimental results. The classification performance of the model is evaluated by comparing the accuracy (ACC), sensitivity (SEN), and specificity (SPE), and the mean of the experimental results for the single-center data is also calculated. Accuracy measures the proportion of subjects correctly classified (*i.e.,* actual ASD is classified as ASD and actual healthy is classified as healthy). Sensitivity represents the proportion of actual ASD subjects correctly classified as ASD, and specificity measures the proportion of actual healthy subjects classified as healthy. The running time means the training time and the inference time.

Experiments settings

The classification accuracy of our proposed model may be affected by a variety of parameters, including: (1) the choice of hypergraph parameters when constructing high-order brain functional connectivity networks, (2) the number of layers of autoencoders in the initial feature selection process, (3) the number of BLS nodes and the window size of each layer. The hypergraph parameters of the model include the nearest neighbor size K obtained based on the hypergraph similarity matrix. The number of autoencoder layers L, the number of nodes in the enhancement layer E, the mapping layer M and the window size W of the BLS are adjusted during the experiment. In our experiments, we adjust all free parameters by 10-fold cross-validation on the training set. Taking into account the effect of the hypergraph construction parameters, we optimize K in the range {4, 5, ..., 12}. Since there is difference in the high-order brain features that can be learned by different layers of the autoencoder, we test the effect of different layers of

TABLE 4 Classification performance of ASD identification achieved by six different methods on four datasets (*i.e.*, OHSU, NYU, USM and UCLA) with rs-fMRI data.

Site	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
OHSU	SVM	53.8 ± 5.2	55.1 ± 6.1	48.9 ± 7.2	52.6 ± 6.7
	SVM-ATM	70.9 ± 3.7	69.9 ± 5.6	66.8 ± 4.1	70.1 ± 5.5
	MLP	64.0 ± 4.5	56.5 ± 3.9	61.6 ± 4.2	60.3 ± 4.7
	Autoencoder	74.0 ± 3.5	66.6 ± 2.9	75.5 ± 4.7	71.5 ± 4.1
	BLS	75.5 ± 5.1	66.3 ± 3.8	72.6 ± 4.9	75.3 ± 3.9
	Ours	80.8 ± 3.8	76.9 ± 7.7	84.6 ± 7.7	83.3 ± 8.3
NYU	SVM	57.1 ± 2.5	50.3 ± 3.5	62.2 ± 2.7	57.8 ± 3.9
	SVM-ATM	71.2 ± 5.1	53.3 ± 4.2	81.0 ± 1.2	69.1 ± 6.5
	MLP	64.3 ± 4.2	68.4 ± 3.7	60.6 ± 3.9	57.1 ± 4.3
	Autoencoder	65.7 ± 3.2	68.8 ± 2.6	63.2 ± 2.7	61.1 ± 4.8
	BLS	69.7 ± 3.5	67.4 ± 6.3	71.1 ± 1.1	70.8 ± 1.4
	Ours	71.4 ± 5.7	75.0 ± 1.5	68.4 ± 9.4	66.7 ± 9.8
USM	SVM	64.7 ± 5.1	60.6 ± 1.9	66.9 ± 5.1	60.7 ± 3.9
	SVM-ATM	69.6 ± 4.6	44.3 ± 3.8	68.2 ± 6.3	61.8 ± 4.3
	MLP	64.1 ± 4.1	61.2 ± 3.8	65.4 ± 4.2	62.9 ± 3.8
	Autoencoder	62.5 ± 2.8	60.0 ± 3.2	66.3 ± 4.5	62.5 ± 4.1
	BLS	76.9 ± 3.1	78.5 ± 2.9	79.8 ± 3.9	82.2 ± 3.9
	Ours	81.6 ± 2.9	84.8 ± 4.1	76.0 ± 0.9	86.6 ± 0.3
UCLA	SVM	65.1 ± 5.7	68.3 ± 3.5	60.8 ± 4.7	65.2 ± 3.3
	SVM-ATM	72.2 ± 3.1	73.8 ± 4.1	68.9 ± 3.8	69.2 ± 2.8
	MLP	71.9 ± 3.5	72.7 ± 2.4	64.8 ± 3.1	66.1 ± 3.2
	Autoencoder	57.7 ± 4.6	68.2 ± 4.1	47.4 ± 4.8	58.5 ± 3.9
	BLS	73.2 ± 2.8	76.4 ± 4.5	65.8 ± 4.9	71.6 ± 3.1
	Ours	79.5 ± 3.1	77.8 ± 3.7	75.0 ± 2.3	79.2 ± 2.3

the autoencoder on the experimental results for single-center and multi-center data, respectively. We adjust the number of layers of the autoencoder in the range {1, 2, ..., 6}. In addition, the number of nodes in the mapping layer and enhancement layer of the BLS and the size of the window also have a significant impact on the classification results, so we test the classification performance under different node settings. We finally find that the hypergraph-based network of high-order brain function connections are constructed with K set to be 5. The number of layers of the autoencoder L is set to be 3. For the multi-center data, we set the parameters as $M = 20$, $E = 10$, $W = 100$. For the single-center data, we set the parameters as $M = 200$, $E = 50$. Depending on the optimal parameters chosen, we can obtain the best experimental results.

Classification performance

We compare our proposed method with: (1) support vector machine (SVM) with RBF kernel, (2) random forests, (3) deep neural network (DNN), (4) Autoencoder, four state-of-the-art methods shown in Table 2.

In the first part of the experiment, the results of the multi-center data are shown in Table 2. Experimental results demonstrate that our proposed method considers the efficiency of the model while maintaining accuracy. Moreover, the time required for ten-fold cross-validation was significantly reduced. Based on the experimental results, it is demonstrated that the high-order brain functional connectivity network constructed based on hypergraphs can capture more correlations between brain regions than the traditional lower-order brain functional connectivity network, and thus obtain more discriminative features. We confirm that BLS further learns the features extracted by autoencoder and greatly reduces the training time. Our method achieves 71.8% accuracy on multi-center data. Experiments show that our proposed method outperforms other state-of-the-art algorithms in terms of accuracy and training time.

At the same time, in order to verify the effectiveness of the proposed method on independent single-center data, we further conduct experiments on 17 single-center datasets. Table 3 shows the results of our experiments, which demonstrate that our method achieves better classification results on small sample datasets compared to the existing state-of-the-art methods. In particular, for the USM dataset, the accuracy is as high as 81.6%. The experiments demonstrate the significant superiority of our proposed method on small sample data as well.

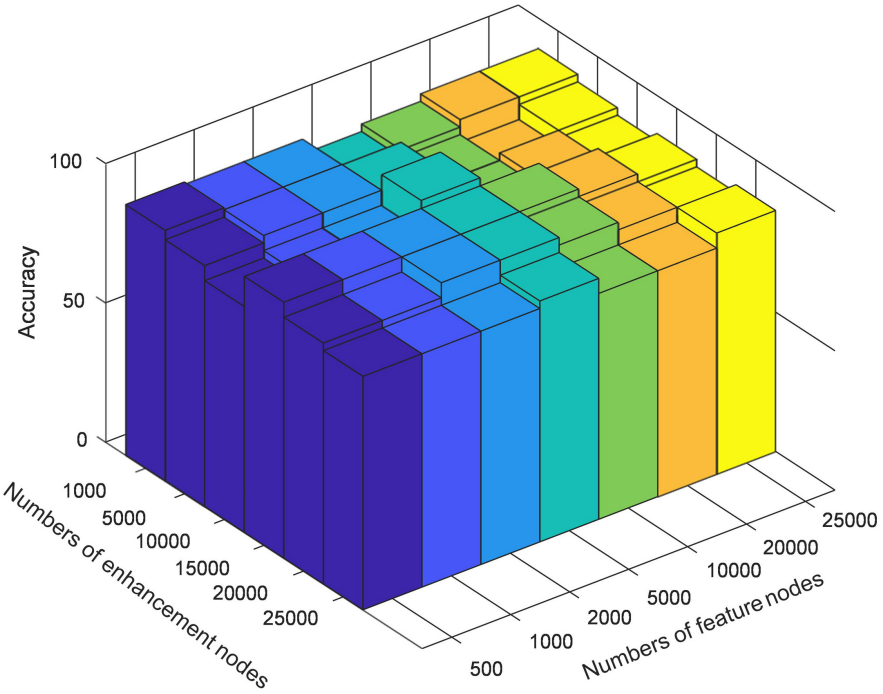


FIGURE 3
The classification accuracy on different mapping features settings.

Discussion

Analysis of the hypergraph learning

To evaluate that BLS plays an important role in our proposed method, we use BLS alone for the final classification task on four representative single-center data. Meanwhile, to demonstrate the improved classification accuracy of high-order brain functional connectivity networks based on hypergraph and BLS, we use BLS to classify the features obtained based on the hypergraph. In particular, we select four representative single-center datasets for comparison based on previous work (Eslami and Saeed, 2019). Our approach is compared with the following methods, as shown in **Table 4**.

(1) We first compare with the SVM method as well as the MLP method. We also compare with SVM methods that incorporated parameter tuning (i.e., SVM as a classifier using the hyperparameter tuning method Auto Tune Models (ATM)) (Eslami and Saeed, 2019).

(2) We then compare with the ASD classification method based on autoencoder and multilayer perceptron proposed by Heinsfeld et al. (2017).

To demonstrate separately that hypergraphs and BLS play an important role in our proposed method, we first used BLS alone to learn the low-order brain functional connectivity network for ASD classification, followed by the construction of a high-order brain functional connectivity network based on hypergraphs, which is then learned and classified by autoencoder and BLS. **Table 4** shows the experimental results of each method, and the results show that our proposed method using only BLS to learn the lower-order brain functional connectivity network significantly outperforms traditional machine learning methods as well as recent deep learning-based methods such as autoencoder-based methods.

The model performance is further improved when the high-order brain functional connectivity network is represented using hypergraphs. Therefore, our proposed method demonstrates significant superiority over other methods.

Analysis of the broad learning system

In order to verify the effectiveness of BLS in further extracting discriminative features of high-order brain functional connectivity networks and optimize the specific structure of BLS, we test the classification results of different mapping features of multi-center data. We empirically set the initial number of enhancement nodes to 1,000, 5,000, 10,000, 15,000, 20,000, 25,000, respectively, and then gradually increase the number of mapping nodes in steps of 500. **Figure 3** shows the variation in model performance at some typical nodes settings during optimization. BLS has obtained classification results with high accuracy under the

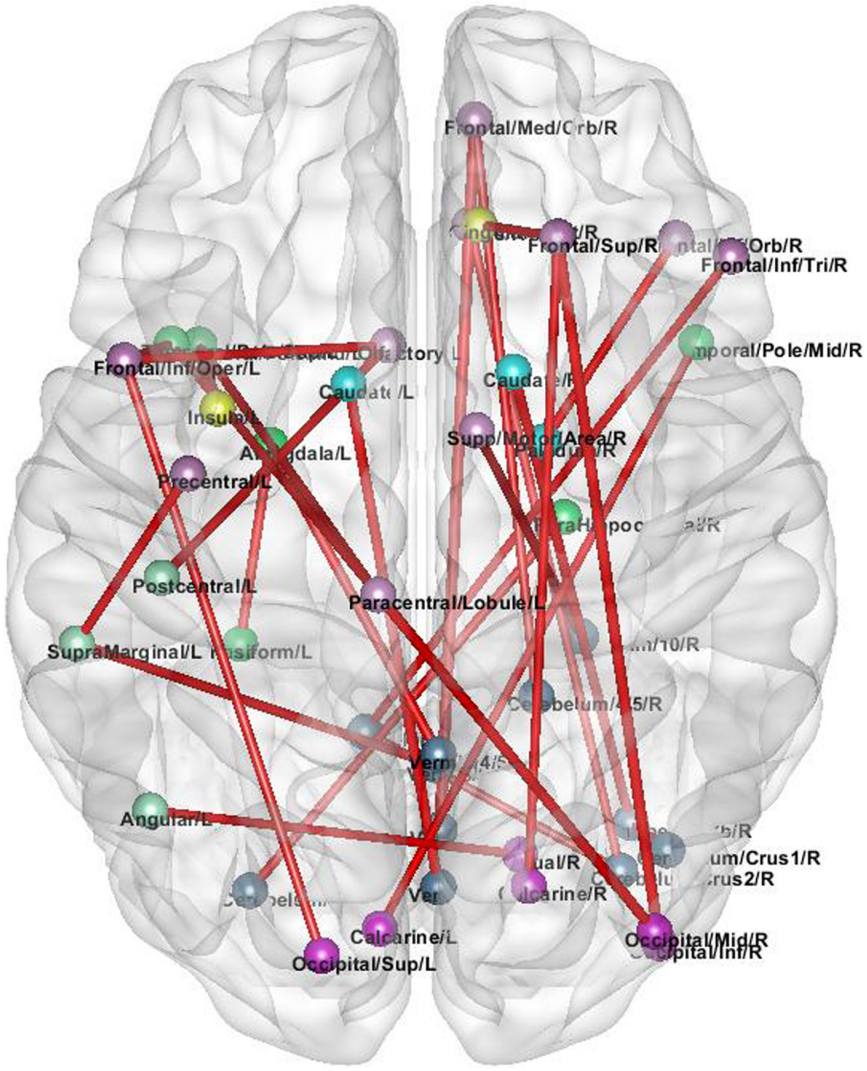


FIGURE 4
The selected high-order brain functional connections.

TABLE 5 Comparison with the state-of-the-art methods using other brain atlas.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
ASD-DiagNet (Eslami et al., 2019)	68.4 ± 1.2	70.9 ± 0.5	65.7 ± 2.6	65.2 ± 3.5
CNN + Element-wise Filters (Xing et al., 2019)	65.7 ± 3.8	70.8 ± 0.1	61.3 ± 7.0	61.3 ± 7.4
Auto-AsD-Network (Eslami and Saeed, 2019)	70.1 ± 1.7	71.7 ± 0.3	68.5 ± 2.8	70.5 ± 5.3
Autoencoder + DNN (Mostafa et al., 2020)	79.1 ± 1.8	77.5 ± 5.8	80.7 ± 12.5	80.0 ± 10.5
Riemannian Regression (Wong et al., 2018)	71.1 ± 1.5	72.7 ± 0.8	69.4 ± 2.9	71.1 ± 4.8
Ours	83.1 ± 3.9	82.2 ± 5.1	80.7 ± 4.2	86.0 ± 6.3

settings of different mapping nodes. In most cases, when the number of enhancement nodes is fixed, model performance becomes better and worse when the number of mapping nodes continuously increases. Therefore, the optimal node setting can be found in this process. When the number of mapping nodes and the number of enhancement nodes reaches 5,000 and 10,000, respectively, the best result, 76.2%, can be obtained.

Analysis of high-order brain connectivity network

In order to explore the high-order connections associated with ASD in the brain functional connectivity network, we analyze the high-order brain functional connections that are selected most frequently. As shown in **Figure 4**. These ROIs are colored according to the static network to which they belong. We find the selected brain connections and regional distribution of brain regions scatter across the two hemispheres and different lobes, showing a pattern of functional abnormalities throughout the brain of ASD patients. Specifically, the top brain FCs are visualized in **Figure 4**. The connections in Red represent the edges of brain ROIs and the ROIs with same color belong to the same brain modules. The brain regions shown in **Figure 4** are highly associated with ASD (Wang et al., 2014). The selected connectivities include the salience network and cerebellum region, and these regions are also shown to be closely related to ASD. These results verify the reliability of our proposed method in detecting informative functional connectivity for ASD identification.

Experiments on other brain atlas

To verify the robustness and scalability of our method on another atlas, we further select the ABIDE dataset for our experiments, using the preprocessing method and brain region segmentation method used by Mostafa et al. (2020) and Yin et al. (2022). In contrast to the aforementioned segmentation of brain regions into 200 ROIs based on the cc200 atlas, we segment the brain into 264 ROIs and then obtain another brain feature representation by calculating the high-order brain functional connectivity network among the 264 ROIs, *i.e.,* we obtain 69,432 pairwise correlation features. We use 871 samples from the ABIDE dataset as in Mostafa et al. (2020); Yin et al. (2022) for our experiments. We compare our proposed method with the latest methods, namely (1) the autoencoder based method for ASD diagnosis proposed by Eslami and Saeed (2019); Eslami et al. (2019), (2) a novel convolutional neural network method proposed by Xing et al. (2019), (3) an autoencoder and DNN classifier based method for ASD diagnosis proposed by Mostafa et al. (2020), and a method based on logarithmic Euclidean and affine invariant Riemann metric connectivity matrices proposed by Wong et al. (2018). **Table 5** shows the algorithm we compared with and the experimental results. Experiments confirm that our proposed method is significantly superior to other methods in characterizing functional connectivity relationships in the brain. Compared to the autoencoder-based methods proposed by Mostafa et al. (2020) and the Euclidean and affine-invariant Riemannian metric-based connectivity matrix-based methods proposed by Wong et al. (2018), our method performs well on another brain atlas. We experimentally demonstrate the robustness and scalability of our method.

Conclusion

We propose a deep-broad learning-based method to explore the high-order brain functional connectivity for ASD diagnosis. Our hypergraph-based higher-order brain functional connectivity network helps to characterize the global features of the brain. The use of autoencoder and BLS to sequentially learn high-order features makes the ASD detection model more efficient and effective. Our experiments are conducted on single-center and multi-center data of the ABIDE dataset. To verify the robustness and scalability of the method, we perform additional experiments on another brain atlas that divide brain regions into 264 ROIs. Experimental results demonstrate that our profiled high-order brain functional connectivity network can represent more discriminative global brain features. The combination of BLS and autoencoder further quickly learns the features, and the diagnostic model can achieve higher accuracy.

Data availability statement

The datasets analyzed for this study can be found in the ABIDE-I repository in Craddock et al. (2013) (http://preprocessed-connectomes-project.org/abide/).

Author contributions

XH: supervision and editing. QA: methodology, writing—original draft, and coding. JL: methodology and coding. HM: writing—original draft. YG: investigation. MY: supervision. JQ: supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62276088, in part

by the Natural Science Foundation of Hebei Province of China under Grant F2020202025, in part by the Project of Strategic Importance Scheme of The Hong Kong Polytechnic University under Grant 1-ZE2Q, and in part by the Hebei Province Research Student Innovation Funding Project under Grant CXZZSS2022037.

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

Agastinose Ronicko, J. F., Thomas, J., Thangavel, P., Koneru, V., Langs, G., and Dauwels, J. (2020). Diagnostic classification of autism using resting-state fMRI data improves with full correlation functional brain connectivity compared to partial correlation. *J. Neurosci. Methods* 345:108884. doi: 10.1016/j.jneumeth.2020.108884

Baggio, H. C., Sala-Llonch, R., Segura, B., Marti, M. J., Valldeoriola, F., Compta, Y., et al. (2014). Functional brain networks and cognitive deficits in Parkinson's disease. *Hum. Brain Mapp.* 35, 4620–4634. doi: 10.1002/hbm.22499

Chen, C. L. P., and Liu, Z. (2017). "Broad learning system: A new learning paradigm and system without going deep," in *Proceedings of the 2017 32nd youth academic annual conference of Chinese association of automation (YAC)* (Hefei: IEEE), 1271–1276. doi: 10.1109/YAC.2017.7967609

Chen, C. L. P., and Liu, Z. (2018). Broad learning system: An effective and efficient incremental learning system without the need for deep architecture. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 10–24. doi: 10.1109/TNNLS.2017.2716952

Chen, C. L. P., Liu, Z., and Feng, S. (2019). Universal approximation capability of broad learning system and its structural variations. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 1191–1204. doi: 10.1109/TNNLS.2018.2866622

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., et al. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* 7:41. doi: 10.3389/conf.fninf.2013.09.00041

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

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. *Proc. Int. Symp. Biomed. Imaging* 2018, 1381–1385. doi: 10.1109/ISBI.2018.8363829

Eslami, T., and Saeed, F. (2019). "Auto-AsD-network: A technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, Niagara Falls, NY, 646–651. doi: 10.1145/3307339.3343482

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., and Saeed, F. (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

Feng, C., Jie, B., Ding, X., Zhang, D., and Liu, M. (2020). "Constructing high-order dynamic functional connectivity networks from resting-state fMRI for brain dementia identification," in *Machine learning in medical imaging*, eds M. Liu, P. Yan, C. Lian, and X. Cao (Cham: Springer), 303–311. doi: 10.1109/TMI.2021.3110829

Gao, Y., Member, S., Zhang, Z., Lin, H., Zhao, X., and Du, S. (2020). Hypergraph learning : Methods and practices. *IEEE Trans. Pattern Anal. Mach. Intell.* 8828, 1–18. doi: 10.1109/TPAMI.2020.3039374

Gong, X., Zhang, T., Chen, C. L. P., and Liu, Z. (2022). Research review for broad learning system: Algorithms, theory, and applications. *IEEE Trans. Cybern.* 52, 8922–8950. doi: 10.1109/TCYB.2021.3061094

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

Hahamy, A., Behrmann, M., and Malach, R. (2015). The idiosyncratic brain: Distortion of spontaneous connectivity patterns in autism spectrum disorder. *Nat. Neurosci.* 18, 302–309. doi: 10.1038/nn.3919

Han, R., Chen, C. L. P., and Liu, Z. (2020). A novel convolutional variation of broad learning system for Alzheimer's disease diagnosis by using MRI images. *IEEE Access* 8, 214646–214657. doi: 10.1109/ACCESS.2020.3040340

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2017). 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

Huang, Z.-A., Zhu, Z., Yau, C. H., and Tan, K. C. (2021). Identifying autism spectrum disorder from resting-state fMRI using deep belief network. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 2847–2861. doi: 10.1109/TNNLS.2020.3007943

Jiang, H., Cao, P., Xu, M. Y., Yang, J., and Zaiane, O. (2020). Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction. *Comput. Biol. Med.* 127:104096. doi: 10.1016/j.compbiomed.2020.104096

Khodatars, M., Shoeibi, A., Sadeghi, D., Ghaasemi, N., Jafari, M., Moridian, P., et al. (2021). Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: A review. *Comput. Biol. Med.* 139:104949. doi: 10.1016/j.compbiomed.2021.104949

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

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

Li, Y., 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

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

Liang, X., Wang, J., Yan, C., Shu, N., Xu, K., Gong, G., et al. (2012). Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: A resting-state functional MRI study. *PLoS One* 7:e32766. doi: 10.1371/journal.pone.0032766

Mostafa, S., Yin, W., and Wu, F.-X. (2020). "Autoencoder based methods for diagnosis of autism spectrum disorder," in *Computational advances in bio and*

medical sciences, eds I. Mândoiu, T. Murali, G. Narasimhan, S. Rajasekaran, P. Skums, and A. Zelikovsky (Cham: Springer), 39–51.

Nickel, R. E., and Huang-Storms, L. (2017). Early Identification of young children with autism spectrum disorder. *Indian J. Pediatr.* 84, 53–60. doi: 10.1007/s12098-015-1894-0

Perkins, E. A., and Berkman, K. A. (2012). Into the unknown: Aging with autism spectrum disorders. *Am. J. Intellect. Dev. Disabil.* 117, 478–496. doi: 10.1352/1944-7558-117.6.478

Schölkopf, B., Platt, J., and Hofmann, T. (2007). "Learning with hypergraphs: Clustering, classification, and embedding," in *Proceedings of the 2006 conference advances in neural information processing systems 19* (Cambridge, MA: MIT Press), 1601–1608.

Wang, M., Zhang, D., Huang, J., Yap, P. T., Shen, D., and Liu, M. (2020). 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, N., Yao, D., Ma, L., and Liu, M. (2022). Multi-site clustering and nested feature extraction for identifying autism spectrum disorder with resting-state fMRI. *Med. Image Anal.* 75:102279. doi: 10.1016/j.media.2021.102279

Wang, S. S.-H., Kloth, A. D., and Badura, A. (2014). The cerebellum, sensitive periods, and autism. *Neuron* 83, 518–532. doi: 10.1016/j.neuron.2014.07.016

Wong, E., Anderson, J. S., Zielinski, B. A., and Fletcher, P. T. (2018). Riemannian regression and classification models of brain networks applied to autism. *Connect Neuroimaging* 11083, 78–87. doi: 10.1007/978-3-030-00755-3_9

Xing, X., Ji, J., and Yao, Y. (2019). "Convolutional neural network with element-wise filters to extract hierarchical topological features for brain networks," in *Proceedings of the 2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, Madrid, 780–783. doi: 10.1109/BIBM.2018.8621472

Yao, D., Sui, J., Wang, M., Yang, E., Jiaerken, Y., Luo, N., et al. (2021). A mutual multi-scale triplet graph convolutional network for classification of brain disorders using functional or structural connectivity. *IEEE Trans. Med. Imaging* 40, 1279–1289. doi: 10.1109/TMI.2021.3051604

Yin, W., Li, L., and Wu, F. X. (2022). A semi-supervised autoencoder for autism disease diagnosis. *Neurocomputing* 483, 140–147. doi: 10.1016/j.neucom.2022.02.017

Yu, R., Zhang, H., An, L., Chen, X., Wei, Z., and Shen, D. (2017). Connectivity strength-weighted sparse group representation-based brain network construction for MCI classification. *Hum. Brain Map.* 38, 2370–2383. doi: 10.1002/hbm.23524

Zhang, F., Wei, Y., Liu, J., Wang, Y., Xi, W., and Pan, Y. (2022). Identification of autism spectrum disorder based on a novel feature selection method and variational autoencoder. *Comput. Biol. Med.* 148:105854. doi: 10.1016/j.compbiomed.2022.105854

Zhang, Y., Zhang, H., Chen, X., Lee, S.-W., and Shen, D. (2017). Hybrid high-order functional connectivity networks using resting-state functional MRI for mild cognitive impairment diagnosis. *Sci. Rep.* 7:6530. doi: 10.1038/s41598-017-06509-0

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 30 November 2022
DOI 10.3389/fnins.2022.1054948

OPEN ACCESS

EDITED BY
Shu Zhang,
Northwestern Polytechnical University, China

REVIEWED BY
Yuan Xue,
Johns Hopkins University, United States
Shiqiang Ma,
Tianjin University, China
Lu Zhang,
University of Texas at Arlington, United States

*CORRESPONDENCE
Senchun Chai
chaisc97@163.com

†These authors have contributed equally to this work and share first authorship

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

RECEIVED 27 September 2022
ACCEPTED 07 November 2022
PUBLISHED 30 November 2022

CITATION
Huang L, Zhu E, Chen L, Wang Z, Chai S and Zhang B (2022) A transformer-based generative adversarial network for brain tumor segmentation.
Front. Neurosci. 16:1054948.
doi: 10.3389/fnins.2022.1054948

A transformer-based generative adversarial network for brain tumor segmentation

Liqun Huang[1†], Enjun Zhu[2†], Long Chen[1], Zhaoyang Wang[1], Senchun Chai[1*] and Baihai Zhang[1]

[1]The School of Automation, Beijing Institute of Technology, Beijing, China, [2]Department of Cardiac Surgery, Beijing Anzhen Hospital, Capital Medical University, Beijing, China

Brain tumor segmentation remains a challenge in medical image segmentation tasks. With the application of transformer in various computer vision tasks, transformer blocks show the capability of learning long-distance dependency in global space, which is complementary to CNNs. In this paper, we proposed a novel transformer-based generative adversarial network to automatically segment brain tumors with multi-modalities MRI. Our architecture consists of a generator and a discriminator, which is trained in min–max game progress. The generator is based on a typical "U-shaped" encoder–decoder architecture, whose bottom layer is composed of transformer blocks with Resnet. Besides, the generator is trained with deep supervision technology. The discriminator we designed is a CNN-based network with multi-scale L_1 loss, which is proved to be effective for medical semantic image segmentation. To validate the effectiveness of our method, we conducted exclusive experiments on BRATS2015 dataset, achieving comparable or better performance than previous state-of-the-art methods. On additional datasets, including BRATS2018 and BRATS2020, experimental results prove that our technique is capable of generalizing successfully.

KEYWORDS

generative adversarial network, transformer, deep learning, automatic segmentation, brain tumor

1. Introduction

Semantic medical image segmentation is an indispensable step in computer-aided diagnosis (Stoitsis et al., 2006; Le, 2017; Razmjooy et al., 2020; Khan et al., 2021). In clinical practice, tumor delineation is usually performed manually or semi-manually, which is time-consuming and labor-intensive. As a result, it is of vital importance to explore automatic volumetric segmentation methods with the help of medical images to accelerate the computer-aided diagnosis. In this paper, we focus on the segmentation of brain tumors with the help of magnetic resonance imaging (MRI) consisting of multi-modality scans. The automatic segmentation of gliomas remains one of the most challenging medical segmentation problems stemming from some aspects, such as arbitrary shape and location, poorly contrasted, and blurred boundary with surrounding issues.

Since the advent of deep learning, Convolutional Neural Networks (CNN) have achieved great success in various computer vision tasks, ranging from classification (LeCun et al., 1998; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; Huang et al., 2017), object detection (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016; He et al., 2017; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020) to segmentation (Chen et al., 2014, 2017; Long et al., 2015; Ronneberger et al., 2015; Lin et al., 2017). Fully Convolution Networks (FCN Long et al., 2015) and in particular "U-shaped" encoder–decoder architectures have realized state-of-the-art results in medical semantic segmentation tasks. U-Net (Ronneberger et al., 2015), which consists of symmetric encoder and decoder, uses the skip connections to merge the extracted features from encoder with those from decoder at different resolutions, aiming at recovering the lost details during downsampling. Owing to the impressive results in plenty of medical applications, U-Net and its variants have become the mainstream architectures in medical semantic segmentation.

In spite of their prevalence, FCN-based approaches are incapable of modeling long-range dependency because of its intrinsic limited receptive field and the locality of convolution operations. Inspired by the great success of transformer-based models in Natural Language Processing (NLP) (Devlin et al., 2018; Radford et al., 2018; Liu et al., 2019; Yang et al., 2019; Clark et al., 2020), a growing number of researchers propose to apply the self-attention mechanism to medical image segmentation, attempting to overcome the limitations brought by the inductive bias of convolution, so as to extract the long-range dependency and context–dependent features. Especially, unlike prior convolution operations, transformers encode a sequence of patches and leverage the power of self-attention modules to pre-train on a large-scale dataset for downstream tasks, like Vision Transformer (ViT) (Dosovitskiy et al., 2020) and its variants.

Simultaneously, for the Transformers applied in medical image segmentation, Generative Adversarial Networks (GAN) has revealed great performance in semantic segmentation. In a typical GAN architecture used for segmentation, GAN consists of two competing networks, a discriminator and a generator. The generator learns the capability of contexture representations, minimizing the distance between prediction and masks, while the discriminator on the contrary maximizes the distance to distinguish the difference between them. The two networks are trained in an alternating fashion to improve the performance of the other. Furthermore, some GAN-based methods like SegAN (Xue et al., 2018) achieve more effective segmentation performance than FCN-based approaches.

In this paper, we explore the integrated performance of transformer and generative adversarial network in segmentation tasks and propose a novel transformer-based generative adversarial network for brain tumor segmentation. Owing to the attention mechanism, transformer has a global receptive field from the very first layer to the last layer, instead of focusing solely on the local information from convolution kernel in each layer, thus contributing to the pixel-level classification and being more suitable for medical segmentation tasks. Besides, CNN learns representative features at different resolutions through cascading relationships, while the attention mechanism pays more attention to the relationship between features, thus transformer-based methods are easily-generalized and not completely dependent on the data itself, such as experiments with incomplete images input in Naseer et al. (2021). Inspired by some attempts (Wang W. et al., 2021; Hatamizadeh et al., 2022) of fusing transformer with 3D CNNs, we design an encoder–decoder generator with deep supervision, where both encoder and decoder are 3D CNNs but the bridge of them is composed of transformer blocks with Resnet. In the contrast of typical "U-shaped" decoder–encoder network, our transformer block is designed to replace the traditional convolution-based bottleneck, for the reason that the self-attention mechanism inside transformer can learn long-range contextual representations while the finite kernel size limits the CNN's capability of learning global information. For pixel-wise brain tumor segmentation task, replacing CNN with transformer blocks on the bottleneck contributes to capturing more features from encoder. Inspired by SegAN (Xue et al., 2018), we adopt the multi-scale L_1 loss to our method with only one generator and one discriminator, measuring the distance of the hierarchical features between generated segmentation and ground truth. Experimental results on BRATS2015 dataset show that our method achieves comparable or better performance than some previous state-of-the-art methods. Compared to existing methods, the main contributions of our approach are listed as follows:

- A novel transformer-based generative adversarial network is proposed to address the brain tumor segmentation task with multi-modalities MRI. To enhance the efficiency of brain tumor segmentation, our method incorporates the concepts of "Transformer" and "Generative adversarial". The generator makes use of the transformer blocks to facilitate the process of learning global contextual representations. As far as we are aware, our work is among the very first ones to explore the combination of transformer and generative adversarial networks and achieve excellent performance in the brain tumor segmentation task.
- Our generator exploits transformer with Resnet module in 3D CNN for segmenting multi-modalities MRI brain tumors. Building upon the encoder–decoder structure, both encoder and decoder in our proposed generator are mainly composed of traditional 3D convolution layers, while the bottom layer of the "U-shaped" structure is a transformer with Resnet module. With Resnet, the

transformer block captures both global and local spatial dependencies effectively, thus preparing embedded features for progressive upsampling to full resolution predicted maps.

- Our loss functions are suitable and effectively applied in generator and discriminator. Adopting the idea of deep supervision (Zhu Q. et al., 2017), we take the output of the last three decoder layers of generator to calculate weighted loss for better gradient propagation. Besides, we leverage a CNN-based discriminator to compute multi-scale L_1 norm distance of hierarchical features extracted from ground truth and segmentation maps, respectively.
- The exclusive experimental results evaluated on BRATS2015 dataset show the effectiveness of each part of our proposed methods, including transformer with Resnet module and loss functions. Comparing to existing methods, the proposed method can obtain significant improvements in brain tumor segmentation. Moreover, our method successfully generalizes in other brain tumor segmentation datasets: BRATS2018 and BRATS2020.

The following outlines the structure of this paper: Section 2 reviews the related work. Section 3 presents the detail of our proposed architecture. Section 4 describes the experimental setup and evaluates the performance of our method. Section 5 summarizes this work.

2. Related works

2.1. Vision transformer

The Transformers were first proposed by Vaswani et al. (2017) on machine translation tasks and achieved a quantity of state-of-the-art results in NLP tasks (Devlin et al., 2018; Radford et al., 2018). Dosovitskiy et al. (2020) then applied Transformers to image classification tasks by directly training a pure Transformer on sequences of image patches as words in NLP, and achieved state-of-the-art benchmarks on the ImageNet dataset. In object detection, Carion et al. (2020) proposed transformer-based DETR, a transformer encoder–decoder architecture, which demonstrated accuracy and run-time performance on par with the highly optimized Faster R-CNN (Ren et al., 2015) on COCO dataset.

Recently, various approaches were proposed to explore the applications of the transformer-based model for semantic segmentation tasks. Chen et al. (2021) proposed TransUNet, which added transformer layers to the encoder to achieve competitive performance for 2D multi-organ medical image segmentation. As for 3D medical image segmentation, Wang W. et al. (2021) exploited Transformer in 3D CNN for segmenting MRI brain tumors and proposed to use a transformer in the bottleneck of "U-shaped" network on BRATS2019 and BRATS2020 datasets. Similarly, Hatamizadeh et al. (2022) proposed an encoder–decoder network named UNETR, which employed transformer modules as the encoder and CNN modules as the decoder, for the brain tumor and spleen volumetric medical image segmentation.

Compared to these approaches above, our method is tailored for 3D segmentation and is based on generative adversarial network. Our generator produces sequences fed into transformer by utilizing a backbone encoder–decoder CNN, where the transformer with Resnet module is placed in the bottleneck. With Resnet, the encoder captures features not only from CNN-based encoder but also from transformer blocks. Moreover, the last three output layers of the encoder are considered to calculate the loss function for better performance. Networks like UNETR employ transformer layers as encoder in low-dimension semantic level, and taking this network as backbone in our method without pre-training easily leads to model collapse during the adversarial training phase. Therefore, we do not choose these networks as our backbone. We find that taking transformer as encoder in low-dimension semantic level needs quantities of pre-training tasks on other datasets to get good results, like TransUNet and UNETR above. As shown in our experiments Section 4.6, transformer-based encoder in low-dimension semantic level performances inferior to CNN-based one when training from scratch. Therefore, we choose to apply transformer only in bottleneck, and remain the low-dimension encode layers as convolutional layers. In this way, we can train from scratch, meanwhile achieving good performance.

2.2. Generative adversarial networks

The GAN (Goodfellow et al., 2014) is originally introduced for image generation (Mirza and Osindero, 2014; Chen et al., 2016; Odena et al., 2017; Zhu J.-Y. et al., 2017), making the core idea of competing training with a generator and a discriminator, respectively, known outside of fixed circle. However, there exists a problem that it is troublesome for the original GAN to remain in a stable state, hence making us cautious to balance the training level of the generator and the discriminator in practice. Arjovsky et al. proposed Wasserstein GAN (WGAN) as a thorough solution to the instability by replacing the Kullback-Leibler (KL) divergence with the Earth Mover (EM) distance.

Various methods (Isola et al., 2017; Han et al., 2018; Xue et al., 2018; Choi et al., 2019; Dong et al., 2019; Oh et al., 2020; Ding et al., 2021; He et al., 2021; Nishio et al., 2021; Wang T. et al., 2021; Zhan et al., 2021; Asis-Cruz et al., 2022) were proposed to explore the possibility of GAN in medical image segmentation. Xue et al. (2018) used U-Net as the generator and proposed a multi-scale L_1 loss to minimize the distance of the feature maps of predictions and masks for the medical image segmentation of brain tumors. Oh et al. (2020) took residual blocks into account under the framework of pix2pix

(Isola et al., 2017) and segmented the white matter in FDG-PET images. Ding et al. (2021) took an encoder–decoder network as the generator and designed a discriminator based on Condition GAN (CGAN) on BRATS2015 dataset, adopting the image labels as the extra input.

Unlike these approaches, our method incorporates the concepts of "Transformer" and "GAN." Our discriminator is based on CNN instead of transformer. In our opinion, owing to the attention mechanism inside transformer, transformer has a more global receptive field than CNN with limited kernel size, thus contributing to pixel-level classification and being more suitable for medical segmentation tasks. However, for image-level medical classification, transformer-based discriminator seems to be less appropriate for its weakness of requiring huge datasets to support pre-training, while CNN is strong enough for classification tasks without pre-training. Motivated by viewpoints above, in our method, the transformer-based generator and CNN-based discriminator are combined to facilitate the progress of segmentation under the supervision of a multi-scale L_1 loss.

3. Materials and methods

3.1. Overall architecture

The overview of our proposed model is presented in Figure 1. Our framework consists of a generator and discriminator for competing training. The generator G is a transformer-based encoder–decoder architecture. Given a multi modalities (T1, T1c, T2, and FLAIR) MRI scan $X \in R^{C\times H\times W\times D}$ with 3D resolution (H, W, D) and C channels, we utilize 3D CNN-based down-sampling encoder to produce high dimension semantic feature maps, and then these semantic information flow to 3D CNN-based up-sampling decoder through the intermediate Transformer block with Resnet (He et al., 2016). With skip connection, the long-range and short-range spatial relations extracted by encoder from each stage flow to the decoder. For deep supervision (Zhu Q. et al., 2017), the output of decoder consists of three parts: the output of last three convolution layers after sigmoid. Inspired by Xue et al. (2018), the discriminator D we used has the similar structure as encoder in G, extracting hierarchical feature maps from ground truth (GT) and prediction separately to compute multi-scale L_1 loss.

3.2. Generator

Encoder is the contracting path which has seven spatial levels. Patches of size 160 × 192 × 160 with four channels are randomly cropped from brain tumor images as input, followed by six down-sampling layers with 3D 3×3×3 convolution (stride = 2). Each convolution operation is followed by an Instance Normalization (IN) layer and a LeakyReLU activation layer.

At the bottom of the encoder, we leverage the transformer with Resnet module to model the long-distance dependency in a global space. The feature maps produced by the encoder is sequenced first and then create the feature embeddings by simply fusing the learnable position embeddings with sequenced feature map by element-wise addition. After the position embeddings, we introduce L transformer layers to extract the long-range dependency and context dependent features. Each transformer layer consists of a Multi-Head Attention (MHA) block after layer normalization (LN) and a feed forward network (FFN) after layer normalization. In attention block, the input sequence is fed into three convolution layers to produce three metrics: queries Q, keys K and values V. To combine the advantages of both CNN and Transformer, we simply short cut the input and output of Transformer block. Thus, as in Vaswani et al. (2017) and Wang W. et al. (2021), given the input X, the output of the transformer with Resnet module Y can be calculated by:

$$Y = x + y_L \tag{1}$$

$$y_i = FFN\left(LN\left(y_i^{'}\right)\right) + y_i^{'} \tag{2}$$

$$y_i^{'} = MHA\left(LN\left(y_{i-1}\right)\right) + y_{i-1} \tag{3}$$

$$MHA\left(Q, K, V\right) = Concat\left(head_1, ..., head_h\right) W^O \tag{4}$$

$$head_i = Attention\left(Q, K, V\right) = softmax\left(QK^T/\sqrt{d_k}\right) V \tag{5}$$

where y_i denotes the output of ith ($i \in [1, 2, ..., L]$) Transformer layer, y_0 denotes X, W^O are projection metrics, d_k denotes the dimension of K.

Unlike the encoder, the decoder uses 3D 2 × 2 × 2 transpose convolution for up-sampling, followed by skip connection and two 3D 3 × 3 × 3 convolution layers. For a better gradient flow and a better supervision performance, a technology called deep supervision is introduced to utilize the last three decoder levels to calculate loss function. Concretely, we downsampled the GT to the same resolution with these outputs, thus making weighted sum of loss functions in different levels.

The detailed structure of our transformer-based generator is presented in Table 1. In the encoder part, patches of size 160 × 192 × 160 voxels with four channels are randomly cropped from the original brain tumor images as input. At each level, there are two successive 3 × 3 × 3 unbiased convolution layers followed by normalization, activation layers and dropout layers. Beginning from the second level, the resolution of the feature maps is reduced by a factor of 2. These features, e.g., areas of white matter, edges of brain, dots and lines, etc., are extracted by sufficient convolution kernels for next blocks. The

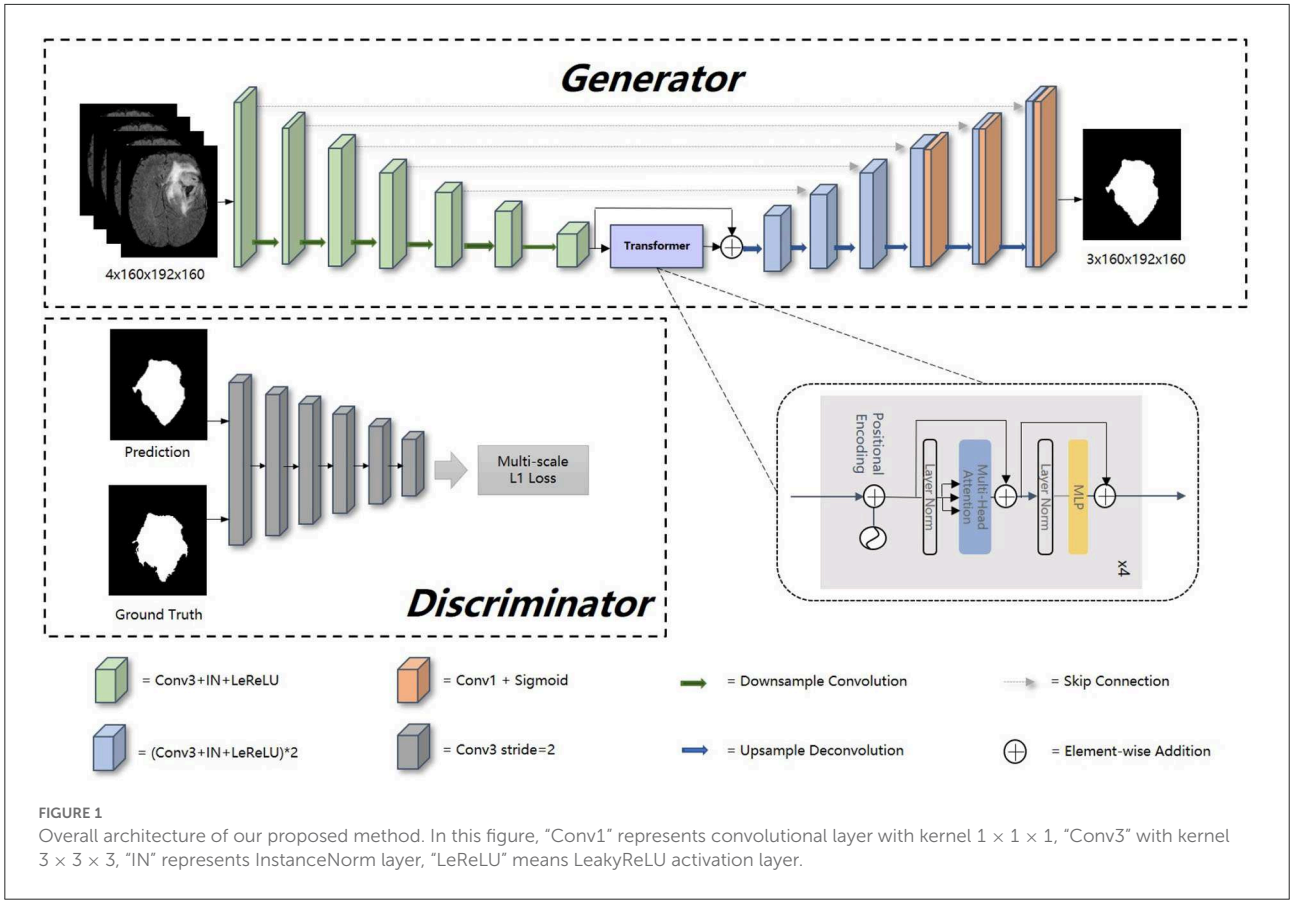


FIGURE 1
Overall architecture of our proposed method. In this figure, "Conv1" represents convolutional layer with kernel $1 \times 1 \times 1$, "Conv3" with kernel $3 \times 3 \times 3$, "IN" represents InstanceNorm layer, "LeReLU" means LeakyReLU activation layer.

transformer block enriches the global contextual representation based on the attention mechanism, forcing features located in the desired regions unchanged while suppressing those in other regions. The shortcut branch crossing the transformer block fusing the features from both encoder part and transformer block by element-wise addition, indicating that our generator is capable of learning short-range and long-range spatial relations with neither extra parameter nor computation complexity. According to the attributes of Resnet (He et al., 2016), $y = f(x) + x$, where $f(x)$ in our method represents transformer blocks, x is the output of CNN-based encoder, whose contexture representations in feature maps are relatively short-range than transformer's. With Resnet, the element-wise addition of $f(x)$ and x can directly fuse the short-range spatial relations from CNN-based encoder and long-range spatial relations from transformer-based bottleneck. Additionally, unlike neural network layers, element-wise addition is a math operation with no more memory cost and negligible computation time cost. The decoder part contains amounts of upsampling layers and skip connection to progressively recover semantic information as well as resolution. The first upsampling layer is implemented by interpolation while the other upsampling layers adapt the form of deconvolution with stride set to 2. At level $i \in [1, 5]$, the encoder block D_i doubles the spatial resolution, followed by skip connection to fuse high-level (from D_i) and low-level (from encoder block E_i) contextual representation so as to segment the desired tumor regions. For a better supervision performance, the outputs of D_i where $i \in [1, 3]$ are fed into $1 \times 1 \times 1$ convolution layer and sigmoid layer to predict segmentation maps with different resolution. Accordingly, the ground truth is downsampled to different shapes such that they match the shapes of those segmentation maps.

Our generator's vital part is the transformer with Resnet module. As shown in Table 1, our transformer with Resnet module consists of transformer block and Resnet, while transformer block is composed of position encodings module, several transformer layers depicted in Figure 2 and features projection module. To make use of the order of the input sequence reshaped from bottom layer feature maps, we introduce a learnable positional encoding vector to represent some information about position of tokens in the sequence, instead of sine and cosine functions. After position encoding and normalization, the input sequence is fed into three different linear layers to create queries, keys, and values. Then, we compute the dot products of keys with queries. To avoid extremely small gradients after softmax function, we scale the dot-products by a factor related to dimensions of queries, as shown in Equation 5. Multiplying scaled weights with

TABLE 1 The detailed structure of proposed generator.

Stage	Name	Details	Output size
Encoder	E1	[Conv3, IN, LeReLU, Dropout]	64*160*192*160
		[Conv3, IN, LeReLU, Dropout]	
	E2	[Conv3(stride2), IN, LeReLU, Dropout]	96*80*96*80
		[Conv3, IN, LeReLU, Dropout]	
	E3	[Conv3(stride2), IN, LeReLU, Dropout]	128*40*48*40
		[Conv3, IN, LeReLU, Dropout]	
	E4	[Conv3(stride2), IN, LeReLU, Dropout]	192*20*24*20
		[Conv3, IN, LeReLU, Dropout]	
	E5	[Conv3(stride2), IN, LeReLU, Dropout]	256*10*12*10
		[Conv3, IN, LeReLU, Dropout]	
	E6	[Conv3(stride2), IN, LeReLU, Dropout]	384*5*6*5
		[Conv3, IN, LeReLU, Dropout]	
	E7	[Conv3(stride2), IN, LeReLU, Dropout]	512*3*3*3
		[Conv3, IN, LeReLU, Dropout]	
Transformer	ResTransBlock	Reshape	512*3*3*3
		PE	
		Transformer Layer*4	
		Reshape	
		Resnet	
Decoder	D6	Upsample	384*5*6*5
		[Conv3, IN, LeReLU, Dropout] x 2	
	D5	Deconv	256*10*12*10
		Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	D4	Deconv	192*20*24*20
		Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
		Deconv	128*40*48*40
	D3	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output3	Conv1 + Sigmoid	4*40*48*40
		Deconv	96*80*96*80
	D2	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output2	Conv1 + Sigmoid	4*80*96*80
		Deconv	64*160*192*160
	D1	Concat	
		[Conv3, IN, LeReLU, Dropout]	
		[Conv3, IN, LeReLU, Dropout]	
	Output1	Conv1 + Sigmoid	3*160*192*160

values, we obtain a single attention output, which is then concatenated with other heads' outputs to produce the multi-head attention outputs. Subsequently, normalization, dropout, and multi-layer perception (MLP) layers are utilized to produce the transformer layer's ultimate output. While convolution layers have local connections, shared weights, and translation equivariance, attention layers are global. We take advantage of both by residual connection to learn both short-range and long-range spatial relations with no more memory cost and negligible computational time cost.

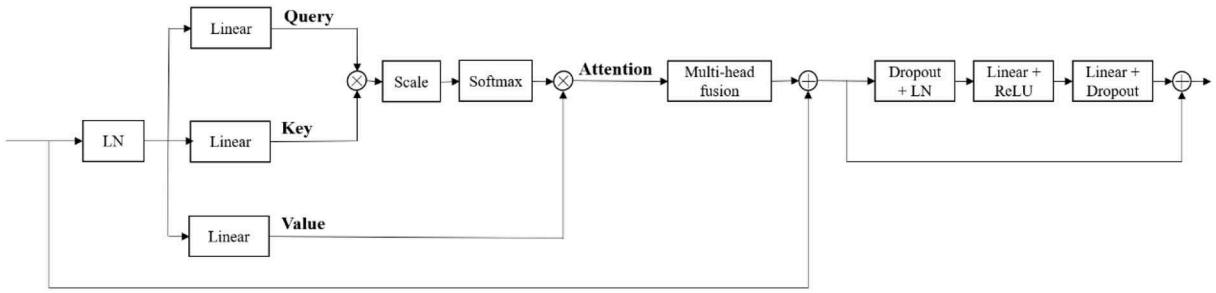


FIGURE 2
The structure of transformer layer.

3.3. Discriminator and loss function

To distinguish the difference between the prediction and GT, the discriminator D extracts features of GT and prediction to calculate L_1 norm distance between them. The discriminator is composed of six similar blocks. Each of these blocks consists of a 3 × 3 × 3 convolution layer with a stride of 2, a batch normalization layer and a LeakyReLU activation layer. Instead of only using the final output of D, we leverage the jth output feature $f_j^i(x)$ extracted by ith ($i \in [1, 2, \ldots, L]$) layers from image x to calculate multi-scale L_1 loss ℓ_D as follows:

$$\ell_D\left(x, x^{'}\right) = \frac{1}{L * M}\sum_{i=1}^{L}\sum_{j=1}^{M}\left\|f_j^i(x) - f_j^i\left(x^{'}\right)\right\|_1 \tag{6}$$

where M denotes the number of extracted features of a layer in D.

Referring to the loss function of GAN (Goodfellow et al., 2014), our loss function of the whole adversarial process is described as follows:

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D}\mathcal{L}\left(\theta_G, \theta_D\right) = \mathbb{E}_{x\sim P_{data}}\left(\ell_D\left(G(x), y\right)\right) \\ +\mathbb{E}_{x\sim P_{data}}\left(\ell_{deep\_bce\_dice}\left(G(x), y\right)\right)\end{aligned} \tag{7}$$

where x, y denote the input image and ground truth, respectively, $\ell_{deep\_bce\_dice}$ denotes that the segmentation maps of generator are used to calculate the BCE loss together with the Dice loss under deep supervision. Concretely, $\ell_{deep\_bce\_dice}$ is a weighted sum of $\ell_{deep\_bce\_dice}\left(p_i, y_i\right), i \in [1, 2, 3]$ for prediction p_i and mask y_i where i denotes the ith level of decoder (D_i).

The detailed training process is presented in Algorithm 1, which interprets the procedure of sampling data and following updating discriminator and generator with corresponding loss function respectively.

for number of training epoches **do**
  for steps of training discriminator **do**
    Get n input images from p_{data} $\{x^1, \ldots, x^n\}$ and corresponding labels $\{y^1, \ldots, y^n\}$.
    Update discriminator by maximizing the loss below:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ell_D\left(G\left(x^i\right), y^i\right)\right]$$

    Clip the weights of discriminator.
  end for
  Get n input images from p_{data} $\{x^1, \ldots, x^n\}$ and corresponding labels $\{y^1, \ldots, y^n\}$.
  Update generator by minimizing the loss below:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\ell_{deep\_bce\_dice}\left(G\left(x^i\right), y^i\right) + \ell_D\left(G\left(x^i\right), y^i\right)\right]$$

end for

Algorithm 1. The detailed training process. $\ell_{deep\_bce\_dice}$ represents BCE Dice loss with deep supervision, ℓ_D represents multi-scale L_1 loss.

4. Experimental results

4.1. Dataset

In the experiments, we evaluated our method using the Brain Tumor Image Segmentation Challenge 2015 (BRATS2015) dataset. In BRATS2015, the training dataset contains manual annotation by clinical experts for 220 patient cases with high-grade glioma (HGG) and 55 patient cases with low-grade glioma (LGG), whereas 110 patient cases are supplied in the online testing dataset without annotation. Four 3D MRI modalities—T1, T1c, T2, and FLAIR—are used for all patient cases, as depicted in Figure 3. Each modality has the origin size 240 × 240 × 155 with the same voxel spacing. The ground truth has five classes: background (label 0), necrosis (label 1), edema (label 2), non-enhancing tumor (label 3), and enhancing tumor (label 4).

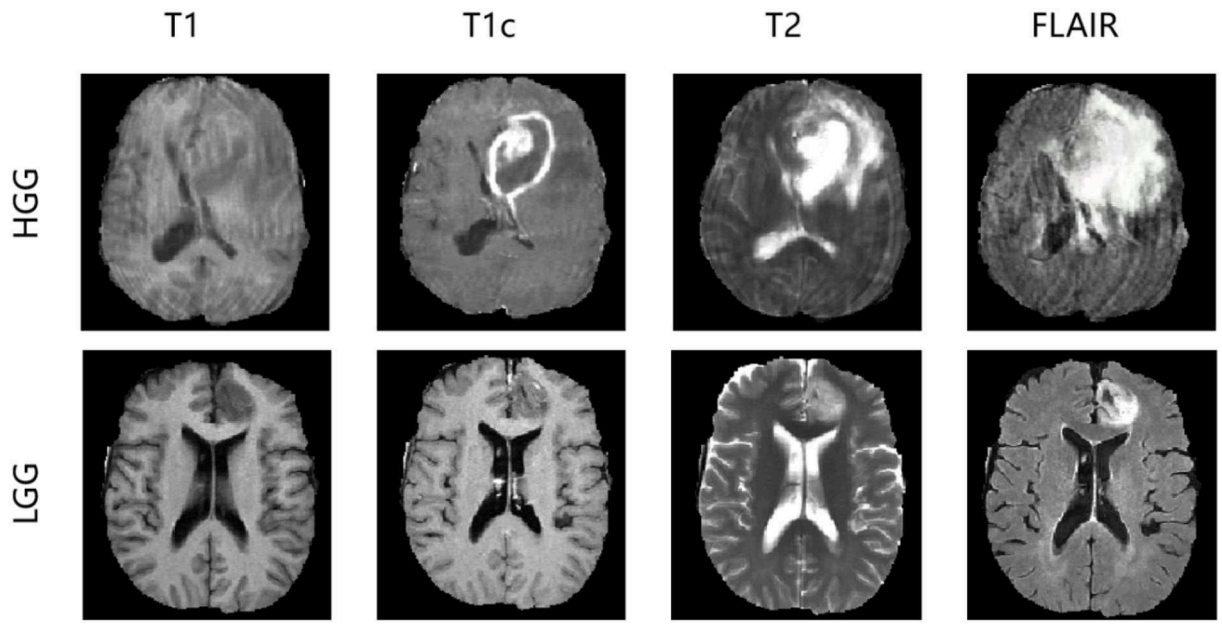


FIGURE 3
HGG and LGG cases with four modalities: T1, T1c, T2, FLAIR on BRATS2015 dataset.

We divided the 275 training cases into a training set and a validation set with the ratio 9:1 both in HGG and LGG. During training and validation, we padded the origin size $240 \times 240 \times 155$ to size $240 \times 240 \times 160$ with zeros and then randomly cropped into size $160 \times 192 \times 160$, which makes sure that the most image content is included.

4.2. Evaluation metric

To evaluate the effectiveness of a segmentation method, the most basic thing is to compare it with the ground truth. In the task of brain tumor segmentation, there are three main evaluation metrics compared with the ground truth: Dice, Positive predictive Value (PPV), and Sensitivity, defined as follows:

$$Dice\,(P,\ T) = \frac{1}{2} \times \frac{\left|P_1 \bigcap T_1\right|}{(|P_1| + |T_1|)} \tag{8}$$

$$PPV\,(P, T) = \frac{\left|P_1 \bigcap T_1\right|}{|P_1|} \tag{9}$$

$$Sensitivity\,(P, T) = \frac{\left|P_0 \bigcap T_0\right|}{|T_0|} \tag{10}$$

where P represents the prediction segmented by our proposed methods, T represents the corresponding ground truth. P_1 and T_1 denote the brain tumor region in P and T, P_0 and T_0 denote the other region except brain tumor in P and T, respectively, $|\cdot|$ calculates the number of voxels inside region, $\cap$ calculates the intersection of two regions. When Dice is larger, PPV and Sensitivity are larger at the same time, the predicted segmentation is considered to be more similar to ground truth, proving that the segmentation method is more effective.

4.3. Implementation details

Experiments were run on NVIDIA A100-PCIE (4×40GB) system for 1,000 epochs (about 3 days) using the Adam optimizer (Kingma and Ba, 2014). The target segmentation maps are reorganized into three tumor subregions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The initial learning rate is 0.0001 and batch size is 4. The data augmentation consists of three parts: (1) padding the data from $240 \times 240 \times 155$ to $240 \times 240 \times 160$ with zeros; (2) randomizing the data's cropping from $240 \times 240 \times 160$ to $160 \times 192 \times 160$; (3) random flipping the data across three axes by a probability with 0.5. Impacted by the volumetric input size, the number of parameters of our network is larger than common 2D networks, generator: 58.0127M, transformer blocks inside generator: 11.3977M, discriminator: 75.4524M. Both the Dice loss in deep supervision and multi-scale L_1 loss are employed to train the network in competing progress. In inference, we converted the transformed three subregions (WT, TC, ET) back to the original labels. Specially, we replace the enhancing tumor with necrosis when the possibility

of enhancing tumor in segmentation map is less than the threshold, which is chosen according to the online testing scores.

4.4. Impact of the number of generators and discriminators

As the BRATS2015 is a multi-label segmentation task, our architecture can be implemented with schemes where the number of generators and discriminators are different. Each implementation scheme in Table 2 is specifically described as follows:

- 1G-1D. The network is composed of one generator and one discriminator. The generator outputs three-channel segmentation maps corresponding to three brain tumor subregions, while the discriminator is fed with three-class masked images concatenated in channel dimension.
- 1G-3D. The network is composed of one generator and three discriminators. The generator outputs three-channel segmentation maps while the discriminators output three one-channel maps, each for one class.
- 3G-3D. The network is composed of three generators and three discriminators. Each generator or discriminator is built for one class. There are three pairs of generators and discriminators, indicating that each pair is trained independently for one class.

4.5. Evaluating the transformer with Resnet module

To evaluate the effectiveness of the transformer with Resnet module, we conduct some ablation experiments. We design the bottom layer of our proposed generator with different schemes as follows:

- Transformer with Resnet. The bottom layer is composed of Transformer with Resnet we proposed.
- Transformer w/o Resnet. The bottom layer is composed of Transformer block, ranging from projection, position embedding to transformer layers, without shortcut crossing them.
- CNN with Resnet. The bottom layer is composed of convolutional layers together with a shortcut crossing them.
- Shortcut. The bottom layer is simply a shortcut connection from the encoder part to the decoder part.

TABLE 2 Results of different number of generators and discriminators.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
1G-3D	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	0.59	**0.90**	**0.73**	**0.73**
1G-1D	0.84	0.72	0.62	0.82	0.78	0.58	0.89	0.72	0.71
3G-3D	0.81	0.68	0.60	0.83	0.74	**0.62**	0.84	0.70	0.63

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 3 Results of different bottom layer in generator.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
Transformer with Resnet	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	0.59	0.90	**0.73**	**0.73**
Transformer w/o Resnet	0.85	0.71	0.61	0.83	0.79	0.60	0.90	0.69	0.68
CNN with Resnet	0.83	0.68	0.58	0.80	0.78	0.58	**0.91**	0.66	0.62
Shortcut	0.82	0.67	0.60	0.82	0.77	**0.63**	0.87	0.67	0.63

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 4 Results of different discriminators training from scratch.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
CNN-based	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	**0.59**	**0.90**	**0.73**	**0.73**
Transformer-based	0.79	0.66	0.58	0.79	0.77	0.55	0.86	0.64	0.66

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 5 Results of different loss function.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
Our method	**0.85**	**0.73**	**0.63**	**0.83**	**0.79**	**0.59**	**0.90**	**0.73**	**0.73**
w/o deep supervision	0.85	0.72	0.61	0.83	0.78	0.57	0.90	0.73	0.71
Single-scale L_1 loss	0.84	0.72	0.61	0.82	0.78	0.58	0.89	0.72	0.71

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

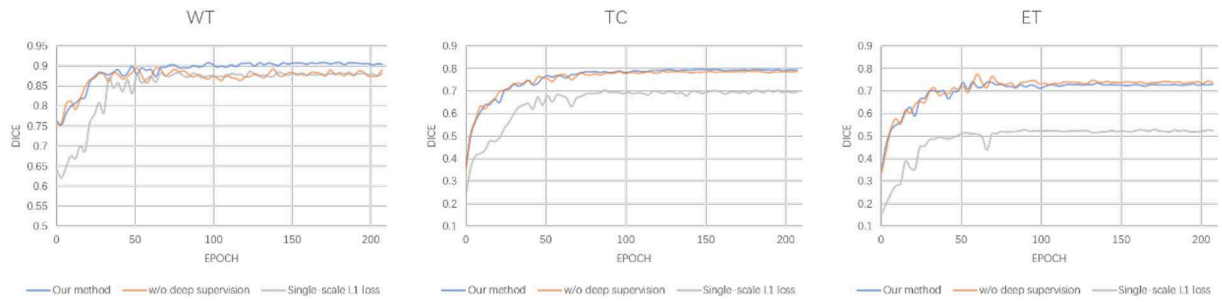


FIGURE 4
Detailed evaluation curves of different loss function.

TABLE 6 Performance of some methods on BRATS2015 testing dataset.

Method	Dice			Positive predictive value			Sensitivity		
	Whole	Core	Enha.	Whole	Core	Enha.	Whole	Core	Enha.
UNET (Ronneberger et al., 2015)	0.80	0.63	0.64	0.83	0.81	**0.78**	0.80	0.58	0.60
ToStaGAN (Ding et al., 2021)	**0.85**	0.71	0.62	0.87	**0.86**	0.63	0.87	0.68	0.69
3D Fusing (Zhao et al., 2018)	0.84	**0.73**	0.62	0.89	0.76	0.63	0.82	**0.76**	0.67
FSENet (Chen et al., 2018)	**0.85**	0.72	0.61	0.86	0.83	0.66	0.85	0.68	0.63
SegAN (Xue et al., 2018)	**0.85**	0.70	**0.66**	**0.92**	0.80	0.69	0.80	0.65	0.62
Our method	**0.85**	**0.73**	0.63	0.83	0.79	0.59	**0.90**	0.73	**0.73**

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

The comparation results are shown in Table 3. From the results, we demonstrate the transformer's superiority and irreplaceability, and we can conclude that transformer with Resnet module make the best of features from transformer block and convolutional encoder to improve the segmentation performance.

4.6. Evaluating the CNN-based discriminator

We select the CNN-based discriminator instead of the transformer-based one as our final discriminator in our architecture, due to our opinion that transformer-based multi-layers discriminator requires huge datasets to support pre-training. To prove that, we conduct ablation experiments to compare their performance by training from scratch. The transformer-based discriminator is implemented using the inspiration of Jiang et al. (2021). Table 4 shows the results on BRATS2015 testing dataset using different discriminators, from which our CNN-based discriminator shows its superior capability of classifying the ground truth and segmentation outputs from scratch. Without pre-training, the CNN-based discriminator appears to be better than the transformer-based one.

4.7. Evaluating the loss function

In this section, we evaluate the effectiveness of the loss function in our proposed methods. As shown in Equation 7, our loss function is divided into two parts: the deep supervision

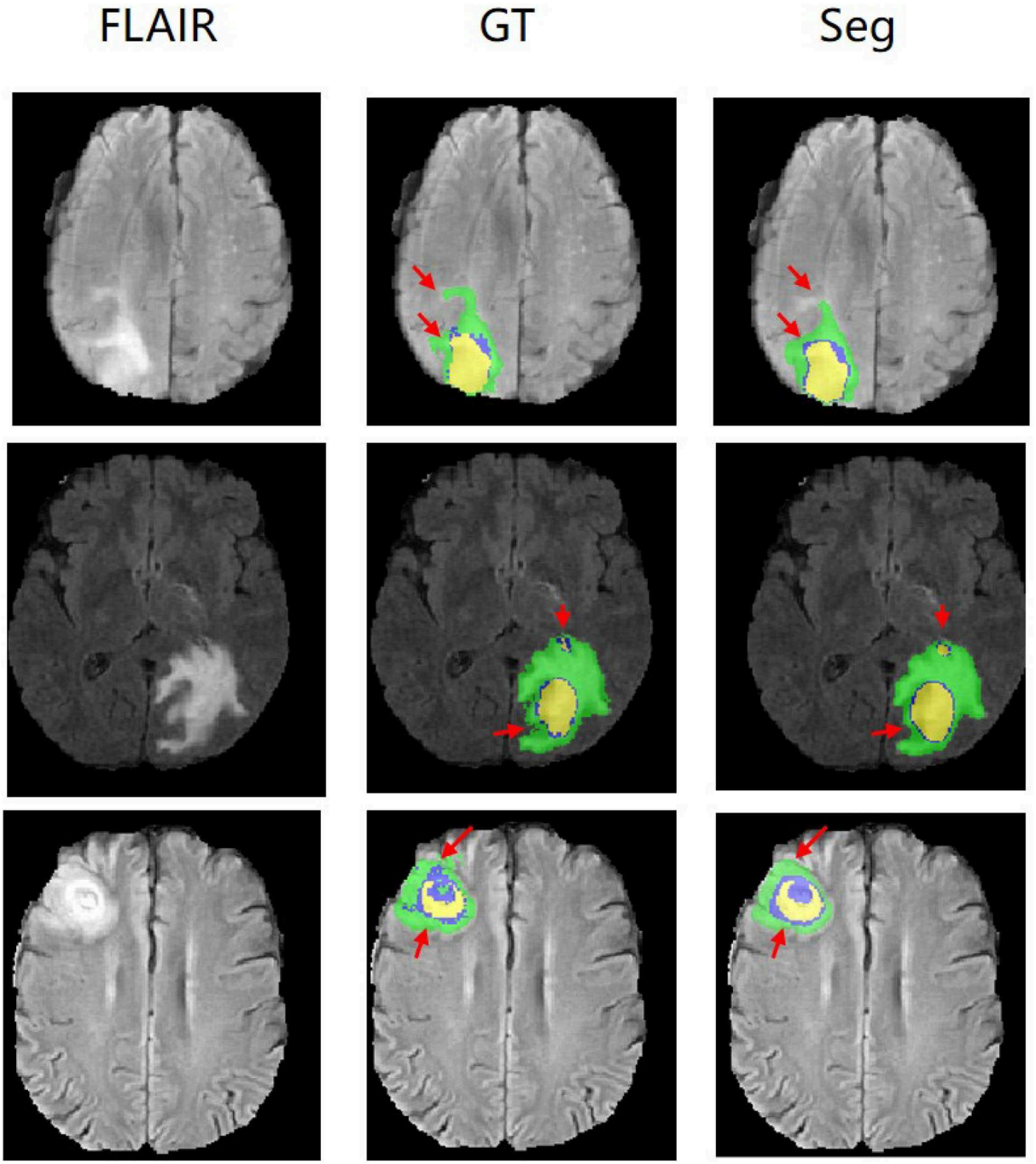


FIGURE 5
Experimental results with corresponding slices on BRATS2015 validation set. The red arrows locate the mainly different regions between GT and segmentation results.

loss and multi-scale L_1 loss. We conduct two ablation experiments: one model with single-scale L_1 loss, the other model without deep supervision loss. It is worth noting that the implementation of these models is the same as 1G-3D where the network consists of one generator and three discriminators and employs the transformer with Resnet module in the bottom layer. From Table 5, we find that our loss function achieves better performance under the same other experimental environment.

The detailed segmentation evaluation scores curves with different loss function are depicted in Figure 4. It is clear that the segmentation performance of all approaches steadily increases

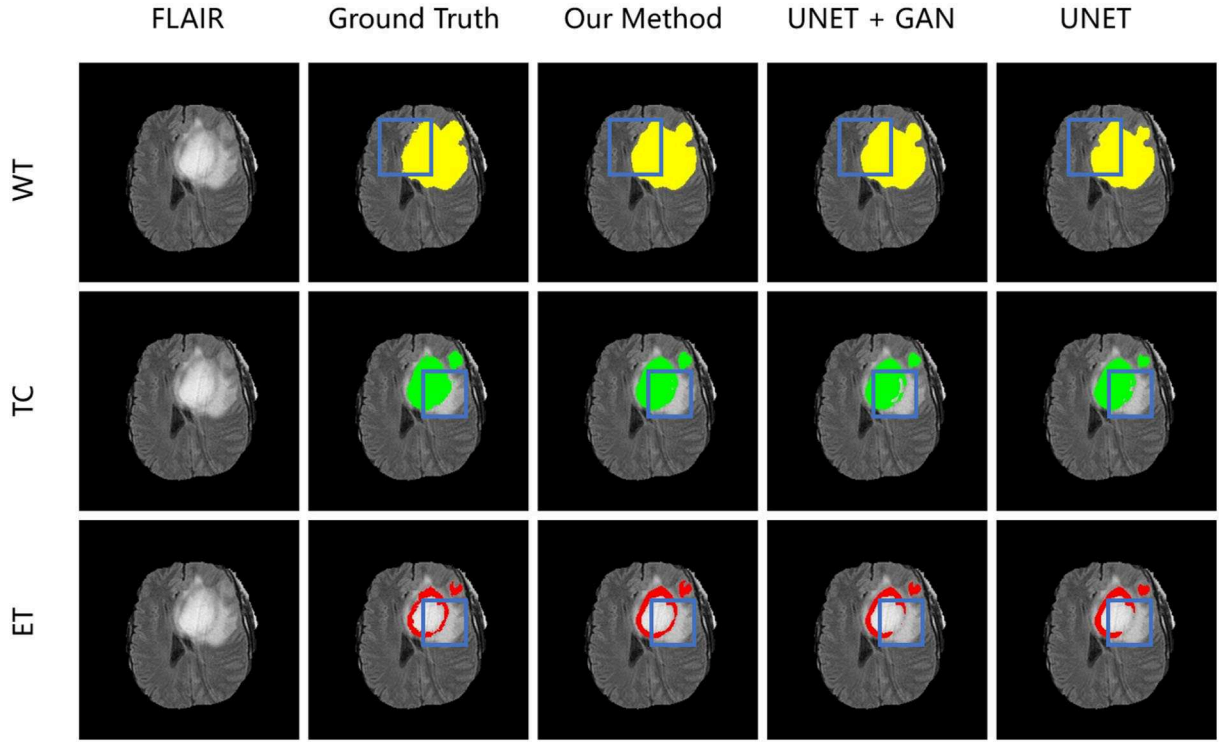


FIGURE 6
From left to right is the visualization results of an original image in FLAIR modality, ground truth, model in our method, model in the form of UNET + GAN, model UNET. From up to down in each column is three segmentation maps predicted with the same method. The blue boxes outline the difference between results from different methods.

as the number of epochs increases until it reaches a steady state. Ranging from WT, TC to ET, our method shows an increasing performance boost over other methods. As a consequence, our method yields the best results in all evaluation metrics listed in Table 5.

4.8. Comparison with other methods

To obtain a more robust prediction, we ensemble 10 models trained with the whole training dataset to average the segmentation probability maps. We upload the results of our methods on the BRATS2015 dataset and get the testing scores computed *via* the online evaluation platform, as listed in Table 6.

Figure 5 shows our qualitative segmentation output on BRATS2015 validation set. This figure illustrates different slices of different patient cases in ground truth and predictions separately.

4.9. Qualitative analysis

To demonstrate the performance of our proposed method, we randomly choose a slice of one patient on BRATS2015 validation set to visualize and compare the result in Figure 6. In Figure 6, images in the same column are produced from the same method, and images in the same row are belonging to the same segmentation label. Concretely, the column FLAIR represents the original image with modality of FLAIR, while other columns are segmentation maps with corresponding categories and colors: WT is yellow, TC is green, and ET is red. The column UNET represents that the corresponding three segmentation maps are inferenced with model UNET. The model of the column UNET plus GAN is built based on UNET, with an addition of GAN, where the generator is UNET with deep supervision and discriminator is a CNN-based network with multi-scale L_1 loss. A deep insight of Figure 6 reveals that with the help of deep supervision and multi-scale L_1 loss, the UNET+GAN method segments fuller edges and richer details than UNET method. When the transformer block is applied, our method produces more smooth borders on the tumor core regions, and more complete contours on enhancing tumor regions. The reason for this improvement seems to be that the transformer with Resnet module can effectively model the short-range and long-range dependency, and collect both local and global contexture representation information. Owing to more complete features, our method achieves the better performance.

TABLE 7 Comparison to other methods on BRATS2018 validation dataset.

Method	Dice(mean)			Hausdorff(mm)		
	Enha.	Whole	Core	Enha.	Whole	Core
Myronenko (2018)	0.7664	0.8836	**0.8154**	3.7731	5.9044	**4.8091**
Hu et al. (2019)	0.7178	0.8824	0.7481	**2.8000**	**4.4800**	7.0700
Chandra et al. (2018)	0.7406	0.8719	0.7990	5.5757	5.0379	9.5884
Liu (2018)	0.7639	0.8958	0.7905	4.0714	4.4924	8.1971
Our method	**0.7686**	**0.9021**	0.8089	5.7116	5.4183	9.4049

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

TABLE 8 Comparison to other methods on BRATS2020 validation dataset.

Method	Dice(mean)			Hausdorff(mm)		
	Enha.	Whole	Core	Enha.	Whole	Core
Tang et al. (2020)	0.703	0.893	0.790	**34.306**	**4.629**	10.071
Zhou et al. (2022)	0.647	0.818	0.759	44.400	10.000	14.600
Anand et al. (2020)	**0.710**	0.880	0.740	38.310	6.880	32.000
Zhang et al. (2021)	0.700	0.880	0.740	38.600	7.000	30.200
Our method	0.708	**0.903**	**0.815**	37.579	4.909	**7.494**

Whole, whole tumor; Core, tumor core; Enha., enhancing tumor.

4.10. Generalization on other datasets

To evaluate generalization of our proposed method, we conduct additional experiments on other datasets relative to brain tumor segmentation, BRATS2018 and BRATS2020, which are composed of more practical patient cases. These datasets differ from BRATS2015 dataset in labels, number of cases and difficulty. The detailed inference performance are listed in Tables 7, 8. On BRATS2018 validation dataset, our proposed method achieves Dice score of 0.7686, 0.9021, and 0.8089, and Hausdorff (HD) of 5.7116, 5.4183, and 9.4049 mm on ET, WT, and TC, respectively. On BRATS2020 validation dataset, our method also realizes Dice score of 0.708, 0.903, and 0.815 and HD of 37.579, 4.909, and 7.494 mm on ET, WT, and TC, respectively. These excellent scores reveal the great generalization of our transformer-based generative adversarial network.

5. Discussion and conclusion

In this paper, we explored the application of a transformer-based generative adversarial network for segmenting 3D MRI brain tumors. Unlike many other encoder–decoder architectures, our generator employs a transformer with Resnet module to effectively model the long-distance dependency in a global space, simultaneously inheriting the advantage of CNNs for learning the capability of local contexture representations. Moreover, the application of deep supervision improves the flowability of gradient to some extent. Our discriminator is applied to measure the norm distance of hierarchical features from predictions and masks. Particularly, we calculate multi-scale L_1 loss between the generator segmentation maps and ground truth. Experimental results on BRATS2015, BRATS2018, and BRATS2020 datasets show a better performance of our proposed method in comparison of other state-of-the-art methods, which proves the superior generalization of our method in brain tumor segmentation.

Data availability statement

The dataset BRATS2015 (Menze et al., 2014; Kistler et al., 2013) for this study can be found in the https://www.smir.ch/BRATS/Start2015. The dataset BRATS2018, BRATS2020 (Menze et al., 2014; Bakas et al., 2017, 2018) and online evaluation platform can be found in this https://ipp.cbica.upenn.edu.

Author contributions

LH: conceptualization, methodology, software, project administration, writing—original draft, writing—review, and editing. EZ: medical expert. LC: validation and project administration. ZW and BZ: supervision. SC: supervision, resources, formal analysis, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was funded in part by the National Natural Science Foundation of China (Grant Nos. 82170374 and 82202139), and also supported in part by the Capital Medical Funds for Health Improvement and Research (CHF2020-1-1053).

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

References

Anand, V. K., Grampurohit, S., Aurangabadkar, P., Kori, A., Khened, M., Bhat, R. S., et al. (2020). "Brain tumor segmentation and survival prediction using automatic hard mining in 3d CNN architecture," in *International MICCAI Brainlesion Workshop* (Virtual: Springer), 310–319.

Asis-Cruz, J. D., Krishnamurthy, D., Jose, C., Cook, K. M., and Limperopoulos, C. (2022). Fetalgan: automated segmentation of fetal functional brain mri using deep generative adversarial learning and multi-scale 3D u-Net. *Front. Neurosci.* 16, 887634. doi: 10.3389/fnins.2022.887634

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data.* 4, 1–13. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Preprint]*. arXiv: 1811.02629. Available online at: https://arxiv.org/pdf/1811.02629.pdf

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. doi: 10.48550/arXiv.2004.10934

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European Conference on Computer Vision* (Glasgow: Springer), 213–229.

Chandra, S., Vakalopoulou, M., Fidon, L., Battistella, E., Estienne, T., Sun, R., et al. (2018). "Context aware 3-D residual networks for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Granada), 74–82.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. doi: 10.48550/arXiv.2102.04306

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*. doi: 10.48550/arXiv.1412.7062

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. *IEEE Trans. Pattern. Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). "Infogan: interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems, Vol. 29* (Barcelona).

Chen, X., Liew, J. H., Xiong, W., Chui, C.-K., and Ong, S.-H. (2018). "Focus, segment and erase: an efficient network for multi-label brain tumor segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 654–669.

Choi, J., Kim, T., and Kim, C. (2019). "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 6830–6840.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*. doi: 10.48550/arXiv.2003.10555

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. doi: 10.48550/arXiv.1810.04805

Ding, Y., Zhang, C., Cao, M., Wang, Y., Chen, D., Zhang, N., et al. (2021). Tostagan: an end-to-end two-stage generative adversarial network for brain tumor segmentation. *Neurocomputing* 462, 141–153. doi: 10.1016/j.neucom.2021.07.066

Dong, X., Lei, Y., Wang, T., Thomas, M., Tang, L., Curran, W. J., et al. (2019). Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Med. Phys.* 46, 2157–2168. doi: 10.1002/mp.13458

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. doi: 10.48550/arXiv.2010.11929

Girshick, R. (2015). "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 580–587.

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, Vol. 27* (Montreal, QC).

Han, Z., Wei, B., Mercado, A., Leung, S., and Li, S. (2018). Spine-gan: Semantic segmentation of multiple spinal structures. *Med. Image Anal.* 50, 23–35. doi: 10.1016/j.media.2018.08.005

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). "UNETR: transformers for 3D medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI: IEEE), 574–584.

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

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.

He, R., Xu, S., Liu, Y., Li, Q., Liu, Y., Zhao, N., et al. (2021). Three-dimensional liver image segmentation using generative adversarial networks based on feature restoration. *Front. Med.* 8, 794969. doi: 10.3389/fmed.2021.794969

Hu, K., Gan, Q., Zhang, Y., Deng, S., Xiao, F., Huang, W., et al. (2019). Brain tumor segmentation using multi-cascaded convolutional neural networks and conditional random field. *IEEE Access* 7, 92615–92629. doi: 10.1109/ACCESS.2019.2927433

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: IEEE), 4700–4708.

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: IEEE), 1125–1134.

Jiang, Y., Chang, S., and Wang, Z. (2021). Transgan: two pure transformers can make one strong gan, and that can scale up. *Adv. Neural Inf. Process. Syst.* 34, 14745–14758. doi: 10.48550/arXiv.2102.07074

Khan, M. Z., Gajendran, M. K., Lee, Y., and Khan, M. A. (2021). Deep neural architectures for medical image semantic segmentation. *IEEE Access* 9, 83002–83024. doi: 10.1109/ACCESS.2021.3086530

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Kistler, M., Bonaretti S., Pfahrer, M., Niklaus, R., and Buchler, P. (2013). The virtual skeleton database: An open access repository for biomedical research and collaboration. *J. Med. Internet Res.* 15, e245. doi: 10.2196/jmir.2930

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems, Vol. 25* (Lake Tahoe).

Le L, Yefeng Z, Gustavo C, Lin Y. (2017). "Deep learning and convolutional neural networks for medical image computing - precision medicine, high performance and large-scale datasets," in *Advances in Computer Vision and Pattern Recognition* (Springer).

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

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). "Refinenet: multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1925–1934.

Liu, M. (2018). "Coarse-to-fine deep convolutional neural networks for multi-modality brain tumor semantic segmentation," in *MICCAI BraTs Conference* (Granada).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: single shot multibox detector," in *European Conference on Computer Vision* (Amsterdam: Springer), 21–37.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. doi: 10.48550/arXiv.1907.11692

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), 3431–3440.

Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imag.* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Mirza, M., and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. doi: 10.48550/arXiv.1411.1784

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 311–320.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. (2021). Intriguing properties of vision transformers. *Adv. Neural Inf. Process. Syst*. 34, 23296–23308. doi: 10.48550/arXiv.2105.10497

Nishio, M., Fujimoto, K., Matsuo, H., Muramatsu, C., Sakamoto, R., and Fujita, H. (2021). Lung cancer segmentation with transfer learning: usefulness of a pretrained model constructed from an artificial dataset generated using a generative adversarial network. *Front. Artif. Intell.* 4, 694815. doi: 10.3389/frai.2021.694815

Odena, A., Olah, C., and Shlens, J. (2017). "Conditional image synthesis with auxiliary classifier gans," in *International Conference on Machine Learning* (Sydney), 2642–2651.

Oh, K. T., Lee, S., Lee, H., Yun, M., and Yoo, S. K. (2020). Semantic segmentation of white matter in fdg-pet using generative adversarial network. *J. Digit. Imaging* 33, 816–825. doi: 10.1007/s10278-020-00321-5

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. (2018). Available online at: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Razmjooy, N., Ashourian, M., Karimifard, M., Estrela, V. V., Loschi, H. J., Do Nascimento, D., et al. (2020). Computer-aided diagnosis of skin cancer: a review. *Curr. Med. Imaging* 16, 781–793. doi: 10.2174/1573405616666200129095242

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 779–788.

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 7263–7271.

Redmon, J., and Farhadi, A. (2018). Yolov3: an incremental improvement. *arXiv preprint arXiv:1804.02767*. doi: 10.48550/arXiv.1804.02767

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

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.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556

Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., and Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. *Nucl. Instrum. Methods Phys. Res. A: Accel. Spectrom. Detect. Assoc. Equip*. 569, 591–595. doi: 10.1016/j.nima.2006.08.134

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 1–9.

Tang, J., Li, T., Shu, H., and Zhu, H. (2020). "Variational-autoencoder regularized 3D multiresunet for the brats 2020 brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Virtual: Springer), 431–440.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*. p. 5998–6008.

Wang, T., Wang, M., Zhu, W., Wang, L., Chen, Z., Peng, Y., et al. (2021). Semi-msst-gan: a semi-supervised segmentation method for corneal ulcer segmentation in slit-lamp images. *Front. Neurosci*. 15, 793377. doi: 10.3389/fnins.2021.793377

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J. (2021). "Transbts: multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Virtual: Springer), 109–119.

Xue, Y., Xu, T., Zhang, H., Long, L. R., and Huang, X. (2018). Segan: adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics* 16, 383–392. doi: 10.1007/s12021-018-9377-x

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). "Xlnet: generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems, Vol. 32* (Vancouver, BC).

Zhan, Q., Liu, Y., Liu, Y., and Hu, W. (2021). Frontal cortex segmentation of brain pet imaging using deep neural networks. *Front. Neurosci*. 15, 796172. doi: 10.3389/fnins.2021.796172

Zhang, W., Yang, G., Huang, H., Yang, W., Xu, X., Liu, Y., et al. (2021). Me-net: multi-encoder net framework for brain tumor segmentation. *Int. J. Imaging Syst. Technol*. 31, 1834–1848. doi: 10.1002/ima.22571

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002

Zhou, J., Ye, J., Liang, Y., Zhao, J., Wu, Y., Luo, S., et al. (2022). scse-nl v-net: A brain tumor automatic segmentation method based on spatial and channel "squeeze-and-excitation" network with non-local block. *Front. Neurosci*. 16, 916818. doi: 10.3389/fnins.2022.916818

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2223–2232.

Zhu, Q., Du, B., Turkbey, B., Choyke, P. L., and Yan, P. (2017). "Deeply-supervised cnn for prostate segmentation," in *2017 International Joint Conference on Neural Networks (IJCNN)* (Anchorage, AK: IEEE), 178–184.

frontiers | Frontiers in Neuroscience

TYPE Hypothesis and Theory
PUBLISHED 01 December 2022
DOI 10.3389/fnins.2022.1058609

OPEN ACCESS

EDITED BY
Dajiang Zhu,
University of Texas at Arlington,
United States

REVIEWED BY
Dalin Yang,
Washington University in St. Louis,
United States
Xiaowei Yu,
University of Texas at Arlington,
United States
Pardis Zarifkar,
Copenhagen University Hospital,
Denmark
Sergio Leonardo Mendes,
Federal University of ABC, Brazil

*CORRESPONDENCE
Yang Liu
liuyang@zust.edu.cn

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

RECEIVED 30 September 2022
ACCEPTED 17 November 2022
PUBLISHED 01 December 2022

CITATION
Guo X, Liu Y, Zhang Y and Wu C (2022) Programming ability prediction: Applying an attention-based convolutional neural network to functional near-infrared spectroscopy analyses of working memory.
Front. Neurosci. 16:1058609.
doi: 10.3389/fnins.2022.1058609

Programming ability prediction: Applying an attention-based convolutional neural network to functional near-infrared spectroscopy analyses of working memory

Xiang Guo[1], Yang Liu[1]*, Yuzhong Zhang[2] and Chennan Wu[1]

[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China, [2]School of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada

Although theoretical studies have suggested that working-memory capacity is crucial for academic achievement, few empirical studies have directly investigated the relationship between working-memory capacity and programming ability, and no direct neural evidence has been reported to support this relationship. The present study aimed to fill this gap in the literature. Using a between-subject design, 17 programming novices and 18 advanced students performed an n-back working-memory task. During the experiment, their prefrontal hemodynamic responses were measured using a 48-channel functional near-infrared spectroscopy (fNIRS) device. The results indicated that the advanced students had a higher working-memory capacity than the novice students, validating the relationship between programming ability and working memory. The analysis results also showed that the hemodynamic responses in the prefrontal cortex can be used to discriminate between novices and advanced students. Additionally, we utilized an attention-based convolutional neural network to analyze the spatial domains of the fNIRS signals and demonstrated that the left prefrontal cortex was more important than other brain regions for programming ability prediction. This result was consistent with the results of statistical analysis, which in turn improved the interpretability of neural networks.

KEYWORDS

programming ability, fNIRS, working memory, convolutional neural network, attention mechanism

Introduction

In the past decade, computer science and programming have been applied in many fields, such as engineering, social sciences music, art, and biology (Buitrago Flórez et al., 2017). Consequently, programming ability has become a basic skill that students may need to master. Several studies have suggested that students with better programming ability have better problem-solving skills and logical reasoning ability (Tu and Johnson, 1990; Shute, 1995; Wing, 2008; Werner et al., 2012; Ivanova et al., 2020; Relkin et al., 2021).

Programming requires memorization of a wide range of information and the ability to manipulate the information at the same time. Students process and hold this information in their working memory, the mode of information storage in the human brain as proposed by cognitive psychology (Baddeley and Hitch, 1974). Working memory is used to store task-relevant information for further application in the process of performing cognitive tasks. It is a memory system with limited capacity for temporary processing and storage of information that supports human thought processes by providing an interface for perception, long-term memory, and action (Baddeley, 2003). Working memory is not only the core of human cognition but also an important component of learning, reasoning, problem-solving, and intellectual activity (Barrouillet and Lépine, 2005; Baddeley, 2010). Working memory plays a critical role in learning. Extensive research has demonstrated a significant relationship between working-memory capacity and academic achievement (Swanson and Alloway, 2012; Anmarkrud et al., 2019). Studies have shown that high performance in math and readings are linked to high working-memory performance (Purpura and Ganley, 2014; Cantin et al., 2016). However, to the best of our knowledge, none of the prior studies showed that programming ability was related to working memory. Nevertheless, since code comprehension involves diverse cognitive domains, including math, logic, and language (Ivanova et al., 2020), programming ability may also be assumed to be related to students' working memory.

The n-back task is one of the most popular experimental paradigms for measuring working memory. The n-back paradigm is a continuous task paradigm (Cohen et al., 1997). In the n-back experiment, participants are asked to monitor a series of verbal/non-verbal stimuli and indicate whether the stimuli currently presented are the same as those that appeared in n trials previously (Braver et al., 1997). The traditional n-back experimental measurements include evaluation of accuracy and reaction time.

In recent years, many researchers have combined functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and functional near-infrared spectroscopy (fNIRS) to measure physiological signals in task-evoked experimental processes to obtain the underlying neuroscientific mechanism of working memory (Ragland et al., 2002; Herff et al., 2013; Lv et al., 2014, 2015; Yeung et al., 2021).

In comparison with fMRI, fNIRS requires a small volume and is lightweight and portable while yielding images with a higher temporal resolution. fNIRS also shows a faster spatial response speed than EEG (Ferrari and Quaresima, 2012; Hong and Yaqub, 2019; Quaresima and Ferrari, 2019; Yang et al., 2019).

Functional near-infrared spectroscopy is a neuroimaging technique for measurement of hemodynamic processes in the brain. In this technique, the absorption of infrared light with a wavelength of 650–950 nm passing through the brain tissue is evaluated to monitor the changes in blood oxygen concentration in different brain tissue regions (Pinti et al., 2020) and obtain insights into the same activation patterns as fMRI. Matthes and Gross (1938) first demonstrated the spectroscopic determination of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) in human tissue in the red and near-infrared regions. In 1993, some research groups proved that fNIRS could be used to investigate brain activity non-invasively (Chance et al., 1993; Hoshi and Tamura, 1993; Villringer et al., 1993). Wolf et al. (2002) first used near-infrared spectroscopy and detected significant changes in the local concentrations of HbO and HbR during brain activity. When the brain executes a task, the increased metabolic demands for oxygen and glucose result in an oversupply of local cerebral blood flow (CBF) to satisfy the increased metabolic demand. CBF is regulated by several neurovascular coupling mechanisms. Therefore, the excessive supply of local CBF leads to an increase in HbO concentration and a decrease in HbR concentration (Pinti et al., 2020). Some previous studies have shown that fNIRS is sensitive to load-dependent working-memory changes in activation (Herff et al., 2013; Meidenbauer et al., 2021) and have demonstrated linear increments in HbO concentrations in frontal activation based on n-back levels (Ayaz et al., 2012; Yeung et al., 2021). The results of a meta-analysis of brain imaging data acquired during the n-back task showed that the participants' prefrontal cortex was activated consistently (Nystrom et al., 2000; Owen et al., 2005). Therefore, in this study, we mainly focused on the concentration changes in HbO and HbR in the prefrontal cortex.

Functional near-infrared spectroscopy is also an effective approach to explore the temporal and spatial states of the human brain (Maki et al., 1995). It provides a balance between temporal and spatial resolution in comparison with other neurophysiological modalities, making it a viable option for mental workload estimation (Isbilir et al., 2019). In the present study, we focused on investigating the channel-wise analysis of fNIRS spatial features to explore the most important brain regions for predicting programming ability.

A recurrent neural network (RNN) is usually considered the best neural network structure for time series prediction, but recent studies have shown that a convolutional neural

network (CNN) can perform these predictions comparably not only with greater accuracy but also more easily and clearly (Bai et al., 2018), particularly when there are many similar time series to learn from Chen et al. (2019). Dilated convolutions can make one-dimensional CNNs effectively learn time series dependencies (Yu and Koltun, 2016; Borovykh et al., 2018). In RNNs, the prediction of subsequent time steps must occur after the previous time step has been completed. In contrast, convolutions can be performed in parallel because the filters used in each layer are the same. Therefore, in training and evaluation, CNNs can process long input sequences simultaneously rather than sequentially as with RNNs (Bai et al., 2018). Since fNIRS signals in the n-back task show multiple similar time series, and we aimed to capture features over the global theoretical receptive fields, a CNN was the best choice for the backbone network in the present study. CNNs have been widely used for automatic fNIRS signal analysis (Trakoolwilaiwan et al., 2017; Janani et al., 2020) and have been used to investigate mental workload levels using multichannel fNIRS signals (Ho et al., 2019; Saadati et al., 2020). One previous study employed a CNN to analyze fNIRS features during an n-back task and proved that CNNs can learn features automatically and obtain accurate results (Yang et al., 2020).

False discovery rate (FDR) measurements (Singh and Dan, 2006) and statistical parametric mapping (SPM) (Koh et al., 2007) have been applied for channel-wise analysis for fNIRS signals. However, these statistical analysis methods corrupt the temporal domain information in fNIRS signals. Among deep neural networks, squeeze-and excitation network (SENet).

SENet, NIRSIT, PET. represents the pioneering concept of channel attention (Guo et al., 2022). The traditional pooling layer reduces the feature map, resulting in damage to channel important information. In contrast, a squeeze-and-excitation (SE) block is a type of attention layer that can collect channel important weight in train processing. The SE block can be used to collect global information, capture channel-wise relationships, and incorporate spatial attention into the structure of the CNN (Hu et al., 2020), thereby improving the interpretability of neural networks. Moreover, in comparison with the application of convolution in feature mapping, the computational cost of SE and weighted summation is very low (Guo et al., 2022). However, SENet, which is an advanced, novel channel-attention network, has not been reported for use in fNIRS signal analysis.

To the best of our knowledge, no previous study has directly explored the brain mechanisms underlying the relationship between working memory and programming ability by using fNIRS data. Thus, the first aim of the current study was to validate the relationship between programming ability and working memory by using an n-back task. The second aim was to investigate whether the fNIRS signals detected during the performance of n-back tasks can predict the participant's programming ability. The third aim was to explore the capability of the attention-based CNN method to analyze the spatial information of the fNIRS signals to identify the optimal brain regions to predict programming ability. Thus, we aimed to explore whether general psychological experiments could be used to predict learners' programming ability, and to provide neuroscience evidence for the findings.

Materials and methods

Participants

Thirty-five participants (17 novices and 18 advanced students) were recruited from the School of Information and Electronic Engineering, Zhejiang University of Science and Technology in China. The novice group included 13 male and four female participants, while the advanced group included 14 male and four female participants. All participants were over 18 years of age [mean ± standard deviation (SD), 20.61 ± 1.23 years].

The novices were freshmen from C++ courses who had not undergone programming-intensive training previously. On the other hand, the advanced students were from the programming competition team who had at least 2 years of programming-intensive training and had at least received an award in the international collegiate programming contest (We did not investigate the effect of programming training on working memory in the present study, and merely used this approach to select participants). Before the experiment, the participants were asked to complete a programming level test, which consisted of ten items with ten points for each completely correct answer and deductions for incomplete results. The maximum total score in the programming level test was 100. The advanced students had higher scores on the programming level test than the novices (mean ± SD, 83.9 ± 5.96 vs. 50.0 ± 7.71). An independent-sample *t*-test revealed that the scores on the programming level test were significantly different between the two groups [F (1, 34) = 214.07, $p < 0.001$, $\eta_p^2 = 0.87$].

All participants signed the informed consent form before the experiment and received a small gift at the end of the study to thank them for their time and effort.

Experimental setup and tasks

The participants were assigned to two groups: novice and advanced students. The experimental procedures were conducted through computer programs based on E-prime, a general psychological experiment software. Each participant was individually tested in a laboratory environment for approximately 30 min. Before completing the task, participants learned about the experimental procedure. The participants were asked to relax and do nothing as the baseline task, and

measurements obtained during this baseline task were used as the baseline for comparison of fNIRS signals.

The present study employed an n-back task to investigate the participants' working memory. The participants monitored a series of character stimuli and responded whenever a stimulus presented was the same as the one presented n trials previously (Owen et al., 2005). The main n-back task involved 30 blocks, with 10 blocks of each n-back level presented pseudorandomly (Braver et al., 1997). **Figure 1** shows the trial schematic of the n-back task conditions in the present study. For example, in the 2-back task, the third C did not match the first A. However, the fourth B matched the second B, so the participants were required respond positively whenever the character they saw was the same as the one they viewed two characters earlier. fNIRS data were recorded continuously during the entire session.

Functional near-infrared spectroscopy data acquisition

In the present study, we focused on the HbO and HbR concentration changes in the prefrontal cortex. The hemodynamic responses measured in the prefrontal cortex were consistent enough to distinguish three levels of n-back workloads (Owen et al., 2005; Herff et al., 2013).

Functional near-infrared spectroscopy (fNIRS) data were recorded at a sampling rate of 8.13 Hz using a wearable NIRS device, the NIRSIT model from OBELAB (Korea). The NIRIST device has a comprehensive 48-channel system and can capture depth-dependent hemodynamic changes in the prefrontal cortex. This system utilizes 24 laser sources (780/850 nm; maximum power under 1 mw) and 32 photodetectors (Choi et al., 2016). The grouping of NIRSIT channels is shown in **Figure 2** and includes the right (#1–16), center (#17–32), and the left (#33–48) regions.

Functional near-infrared spectroscopy data pre-processing

Figure 3 presents the flow diagram for fNIRS data pre-processing. We first loaded fNIRS data into the NIRSIT Analysis Tool for visual inspection, segmentation of the main n-back trials from practice trials, and division of the prefrontal cortex into three regions—right, center, and left—as shown in **Figure 2**. We then performed visual inspection at the participant level to examine overall data quality and to evaluate the quality of the data obtained from both sides of the prefrontal cortex, which showed a much lower signal-to-noise ratio (SNR) than the data from the center of the prefrontal cortex.

Since each participant had a different completion time, we tailored the data with the shortest completion time. Thus, the three n-back tasks had different data lengths. However, in

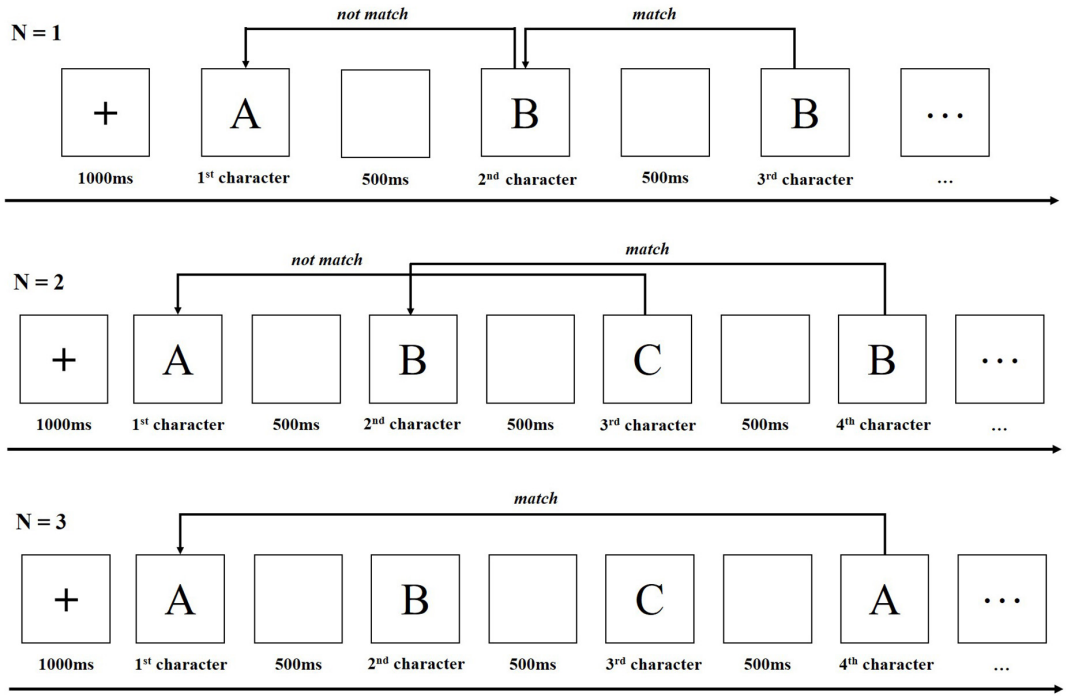


FIGURE 1
Trial schematic of the n-back task conditions.

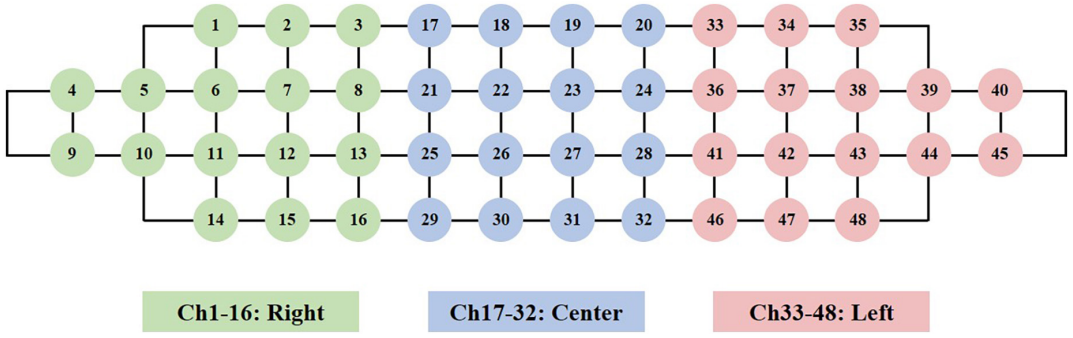


FIGURE 2
Channel configuration in our experiment.

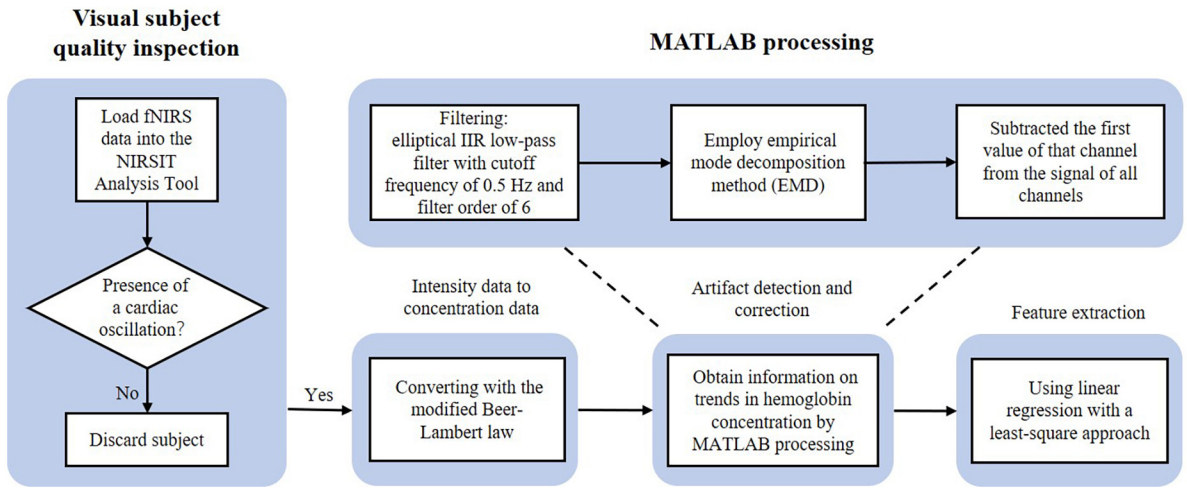


FIGURE 3
Flow diagram of functional near-infrared spectroscopy (fNIRS) data pre-processing for the study.

the subsequent data processing, we focused on the slope and statistical features that are less affected by data length.

Visual inspection was performed by examining the spectrogram of every channel to identify the presence of cardiac oscillations, which are typically approximately 1 Hz (Tong et al., 2011). The presence of this cardiac signal is a good indicator that the optical density signals are successfully coupled with a physiological hemodynamic response (Hocke et al., 2018). This method was employed for preliminary selection of participants. In this visual inspection, one participant with unusable data, which was defined by the presence of more than seven unclean channels in one area, was identified and excluded from further analyses. Thus, the novice and advanced student groups included data from 17 participants each.

Then, we used the NIRSIT Analysis Tool to convert the raw light intensity data into HbO and HbR concentrations by means of the modified Beer–Lambert law (Sassaroli and Fantini, 2004). However, the signals still contained biological and technical artifacts. Several cardiovascular phenomena, such as heart beats, respiration, and blood pressure (Mayer waves), influenced the recorded data. Movement artifacts such as high-frequency spikes, shifts from baseline intensity, and low-frequency variations, which are present in most fNIRS datasets, can severely affect the quality of recorded data (Franceschini et al., 2006; Huppert et al., 2009).

Therefore, we conducted further data processing in Python-SciPy. To attenuate heartbeat and other biological signals, we used an elliptical Infinite Impulse Response (IIR) low-pass filter with a cutoff frequency of 0.5 Hz and a filter order of 6, which robustly removed biological artifacts in the data (Herff et al., 2013). Then, we tried to use the wavelet artifact removal method to reduce the effect of movement artifacts. Since the signals showed channel- and participant-wise variations and the wavelet basis function had limited adaptability, we found it difficult to identify a suitable wavelet basis function to remove the movement artifacts effectively.

Therefore, we used the empirical mode decomposition (EMD) method, which can decompose signals without any additional parameters and therefore robustly reduced the influence of movement artifacts and Mayer wave-like effects in the data. Some of the spontaneous physiological information, such as breathing rate (∼0.3 Hz) and very low-frequency oscillations (<0.01 Hz), were still reflected in the data obtained for post-processing analysis. Therefore, we used the EMD method to deal with this noise after applying a low-pass filter, since direct application of a high-pass filter would have destroyed other useful signal components. Furthermore, due to various factors, the amplitude and intensities of the acquired hemodynamic signals differed significantly among participants. To attenuate the influence of a single participant's data on the grand average of acquired hemodynamic signals, we subtracted the first value of that channel from the signal of all channels and rejected channels that may have significantly influenced the grand average. After these treatments, we obtained information on trends in HbO and HbR concentrations.

The slope of the trend data (Herff et al., 2013) is often used as a simple but effective feature. To obtain this slope, we fitted a straight line to the data using linear regression with a least-squares approach.

Attention-based convolutional neural network for functional near-infrared spectroscopy spatial feature analysis

The structure of the attention-based CNN that we introduced to analyze the fNIRS signals is depicted in **Figure 4**, with the channel-attention blocks showing the spatial importance of fNIRS channels. After data pre-processing as described in Section "Functional near-infrared spectroscopy data pre-processing," each channel of data was resampled and rescaled to a uniform length $L = 256$. Then, we stacked all channels to build a 48×256 feature matrix and used direct resampling and rescaling. The fNIRS signals did not contain periodic frequency information, which may have been corrupted by those processes.

In our training procedure, the mean-squared-error (MSE) method was chosen for the loss function.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

Here, we choose Adam as our optimizer and set the initial learning rate to 0.01.

Three SE blocks were inserted into three normal convolution layers. After the global average pooling operator, each channel data point was consolidated into one data point. In the first SE block, the fully connected layer transforms the 1×48 vector to 1×24; this process is also called the "squeeze." The squeezing function also serves to embed the global distribution of feature responses over all channels. This operator is followed by an excitation operator, which consists of a fully connected layer and a sigmoid activation layer. Excitation is a self-gating mechanism (Hu et al., 2020) that produces a mask representing the per-channel modulation weights. These weights are then applied to the original feature map to generate the new output. This series operation is also known as the self-attention operation (Vaswani et al., 2017). The feature vector needs to be squeezed small and then return to the origin scale *via* excitation because we aimed to improve the training pressure and to prevent overfitting. The reduction factor needs to be carefully adjusted within the training process. A special classified task was chosen. After proper training, we opened the SE block to observe the channel-wise weights mask.

The backbone network of the attention-based CNN we used (as shown in **Figure 4**) had a traditional CNN structure. Here, FC refers to the fully connected layer; BN refers to the batch-normalization layer; and the two round circles indicate the dot-product operator. However, to prevent mixing of the channel information, a generic convolution (GC) layer cannot be used at the beginning of the network. As illustrated in **Figure 4**, the key point is to replace the first GC layer with a depth-wise convolution (DC) layer before the self-attention mechanism finds the important channels. Nevertheless, the other convolution layers are still GC layers. The DC layer ignores the interchannel information, which must be remedied with a point-wise convolution (PC) layer. This will complicate the overall structure of our neural network.

Due to the limited amount of training data in this study, the neural network was easily overfitted. However, in our training procedure, we were not overtly concerned with the generalization properties of the neural network. Instead, we aimed to reveal the importance of channels. The main purpose of this model is to train the channel-attention block, and some degree of overfitting can help make important channels more obvious (Hu et al., 2020). Thus, model overfitting can be acceptable for general inference.

Results

The criterion for statistical significance was set at $p < 0.05$. The Greenhouse–Geisser correction was used to compensate for sphericity violations. Effect sizes were measured by η_p^2, with $\eta_p^2 = 0.01$, 0.06, and 0.14 indicating small, medium, and large effects, respectively (Fritz et al., 2012).

n-back performance

The descriptive statistics for accuracy and reaction time in each group are presented in **Table 1**. Accuracy was calculated by determining the average percentage of correct trials under

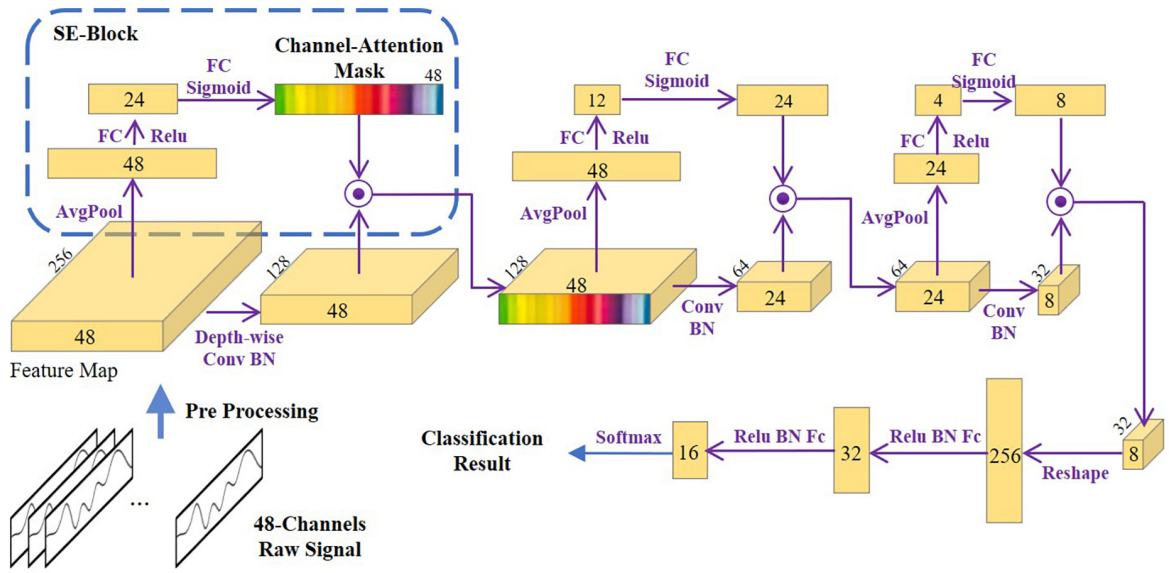


FIGURE 4
The structure of the attention-based convolutional neural network.

each back condition, while reaction time was computed by determining the mean across correct trials for each back condition. As shown in Table 1, novices performed the trials with lower accuracy and slower reaction times than advanced students over each back level.

Paired *t*-tests indicated that the accuracies for the 1-back and 2-back conditions were near the ceiling levels and did not differ in both advanced students and novices. Accuracy for the 3-back condition was marginally significantly lower than those for the 1-back [t (16) = 1.984, p = 0.073] and the 2-back [t (16) = 2.072, p = 0.063] conditions among the advanced students. However, accuracy for the 3-back condition was significantly lower than those for the 1-back [t (16) = 4.690, p = 0.001] and the 2-back [t (16) = 3.801, p = 0.003] conditions among the novices.

Paired *t*-tests indicated that the 1-back task was performed faster than the 2-back [t (16) = −4.641, p = 0.001; t (16) = −4.935, p < 0.001] and 3-back [t (16) = −8.567, p < 0.001; t (16) = −11.950, p < 0.001] tasks by the advanced students and the novices, respectively, while the 2-back task was performed faster than the 3-back [t (16) = −7.637, p < 0.001; t (16) = −8.368, p < 0.001] task by both advanced students and the novices.

To examine group differences, we conducted one-way repeated-measures analysis of variance (ANOVA) with programming ability (novices vs. advanced students) as the between-subjects factor and n-back levels (1-back, 2-back, and 3-back) as the within-subject factor.

One-way repeated-measures ANOVA revealed a marginally significant main effect of programming ability on accuracy [F (1, 32) = 3.867, p = 0.075, $\eta_p^2 = 0.260$]. The interaction between programming ability and the n-back task was not significant [F (2, 64) = 1.692, p = 0.207, $\eta_p^2 = 0.133$].

One-way repeated-measures ANOVA with Greenhouse–Geisser correction revealed a main effect of programming ability on reaction time [F (1, 32) = 5.650, p = 0.029, $\eta_p^2 = 0.239$]. The interaction between programming ability and the n-back task was also not significant [F (2, 64) = 1.177, p = 0.304, $\eta_p^2 = 0.061$].

Figure 5 illustrated that the correlations between reaction time and programming score were negative for 1-, 2-, and 3-back levels (r = −0.75, r = −0.71, and r = −0.81), i.e., the reaction time was faster for a higher programming score.

Functional near-infrared spectroscopy hemodynamic responses

To determine the differences in hemodynamic responses between novices and advanced students under the three n-back conditions, we first analyzed the grand averages of all participants.

Figure 6 exhibited the grand averages of all participants for the three n-back levels. The blue lines showed the grand averages for novices, while the magenta line showed the overall mean for advanced students. For HbO, a clear increase was observed at the 1-, 2-, and 3-back levels, and the slope was positive for all three n-back conditions in the left, center, and right prefrontal cortices. The grand average increase was steeper in the 2-back task than in the 1-back task and was the steepest in the 3-back task.

For HbR, a slight decrease in concentration changes can be seen for all three n-back conditions, and the slope was negative

TABLE 1 Means and standard deviations of accuracy and reaction time during the n-back task.

Dependent variable	Accuracy				Reaction time (ms)			
	Novices		Advanced students		Novices		Advanced students	
	M	SD	M	SD	M	SD	M	SD
1-back	0.986	0.023	0.992	0.019	550	113	474	53.2
2-back	0.983	0.033	0.986	0.032	751	152	654	75.5
3-back	0.845	0.080	0.931	0.065	1276	222	1087	196

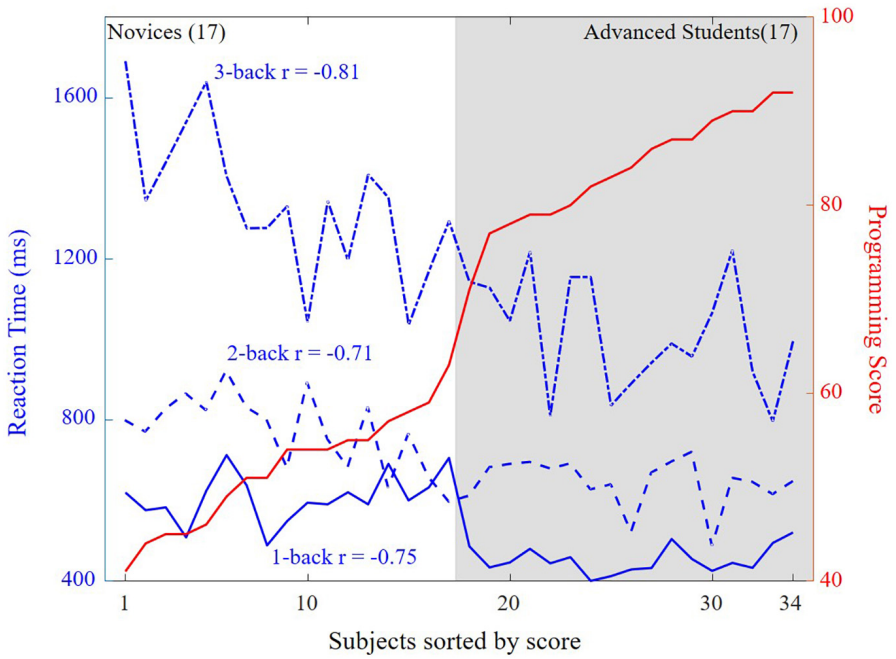


FIGURE 5
Correlations between reaction time and programming score for the three n-back levels.

for the three n-back levels; conversely, there was no obvious difference between the 1- and 2-back grand averages and the 3-back grand average.

One-way repeated-measures ANOVA revealed a main effect of programming ability on HbO [F (1, 32) = 8.838, p = 0.007, η_p^2 = 0.287; F (1, 32) = 12.713, p = 0.002, η_p^2 = 0.366; F (1, 32) = 25.805, p < 0.001, η_p^2 = 0.540] in the right, center, and left prefrontal cortices. The interaction between programming ability and the n-back task was not significant.

Figure 7 illustrated that the correlations between HbO and programming score were negative for the three n-back levels in the left (r = −0.50, −0.51, −0.54), center (r = −0.36, −0.17, −0.38), and right (r = −0.09, r = −0.23, r = −0.38) prefrontal cortices, i.e., HbO is lower for higher programming scores.

One-way repeated-measures ANOVA revealed that the main effect of programming ability on HbR [F (1, 32) = 0.001, p = 0.980, η_p^2 < 0.001; F (1, 32) = 2.716, p = 0.114, η_p^2 = 0.110; F (1, 32) = 0.104, p = 0.750, η_p^2 = 0.005] was not significant in the right, center, or left prefrontal cortices, and there was no interaction between programming ability and the n-back task.

The results indicated that HbO can indicate working-memory load and show significant associations between brain activity and programming ability, but HbR cannot.

Functional near-infrared spectroscopy feature analysis using attention-based convolutional neural network

To obtain the most important channels in the fNIRS signals, we constructed a virtual classification task, and tried letting the neural network model illustrate the importance of the fNIRS channels through the virtual training task. Under this virtual task, we directly combined all subject data into a signal batch

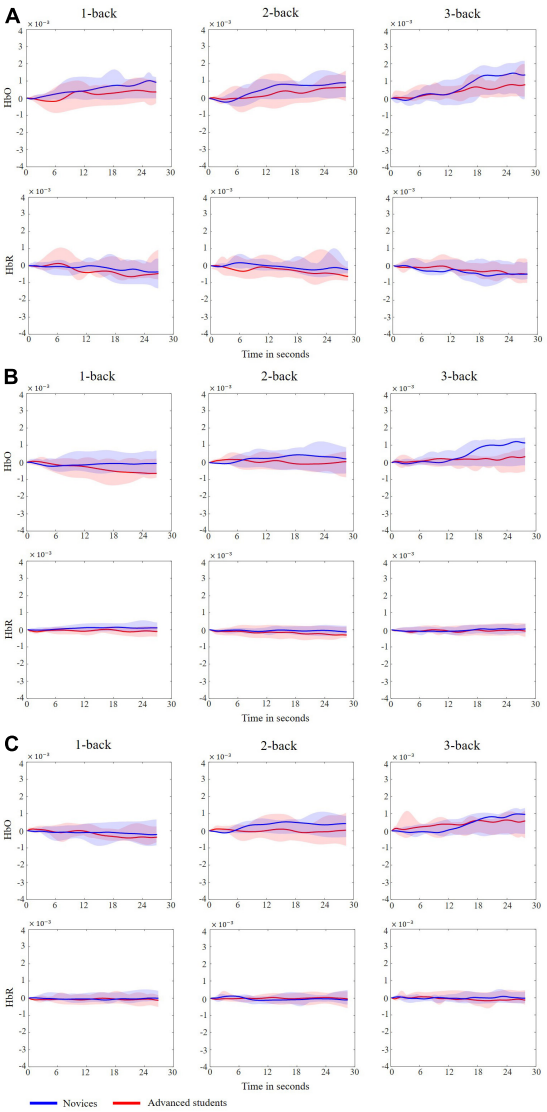


FIGURE 6
Grand averages of all participants in the three n-back levels. **(A)** Left prefrontal cortex. **(B)** Center prefrontal cortex. **(C)** Right prefrontal cortex.

and used MSEloss to train the network. After approximately 50 epochs, the loss stopped falling, and we obtained almost 100% accuracy. Because the amount of data was not very large, the generalization ability of the network was weak. However, we were not going to use this trained network for general classification in a new set of data. We only needed to observe the network's understanding of channel importance under this task. In the present study, we used the mean value of 10 training sessions for subsequent analysis.

The left panels in **Figure 8** showed the channel weights fitting with the training process. The left panel showed that the accuracy (red curve) was close to 100% with the loss (blue curve) down to zero. Furthermore, we distinguished those weights into three brain regions (see **Figure 2**), as shown in the right panels. After fitting, the left prefrontal cortex showed an obviously high weight in the three n-back levels. These findings were consistent with the results of one-way repeated-measures ANOVA, in which the left prefrontal cortex had a larger effect size than the right and center prefrontal cortices.

Discussion

The main purpose of the present study was to investigate the brain mechanisms underlying the relationship between working memory and programming ability by using fNIRS signals.

The analysis of participants' n-back performance showed differences in the accuracy and reaction time depending on the n-back level between novices and advanced students. Advanced students performed better than novices in terms of both accuracy and reaction time. Since the n-back task is recognized as an effective method to measure working memory, a better n-back performance indicated higher working-memory capacity (Kirchner, 1958). The study results validated the relationship between programming ability and working memory, and students with higher working-memory capacity showed better programming ability. The results also provided evidence that limited working-memory capacity has negative effects on learning (Alloway, 2009).

Since the n-back task may be easier for advanced students, the advanced participants were expected to show less prefrontal cortex activation during each n-back experiment (Asgher et al., 2019; Khoe et al., 2020). **Figures 6**, **7** illustrated the neural evidence of this finding. The hemodynamic responses of HbO associated with n-back stimulus presentation increased more in novices than in advanced students. The results of the statistical analyses revealed that HbO in the prefrontal cortex showed significant differences between novices and advanced students during the n-back task. Thus, HbO signals measured during the n-back test can be used to robustly predict the programming ability of students. The changes in cerebral blood oxygen signals represent the changes in local oxygen consumption caused by local brain activity and reflect the activity state of the human brain (Strangman et al., 2002). According to the neural efficiency hypothesis (Haier et al., 1988), the higher the performance in related fields (the higher the cognitive ability), the lower the activation degree of the cerebral cortex, showing a negative correlation. Thus, in comparison with a lower cognitive ability group, a higher cognitive ability group shows lower activation of brain regions when performing tasks with the same difficulty (Dunst et al., 2014; Genc et al., 2018). A higher working-memory load tends to produce greater prefrontal cortex activation (Herff et al., 2013). The novices exhibited significantly higher HbO concentration increments than their advanced counterparts during the n-back tests. Thus, the working-memory load in novices was higher and

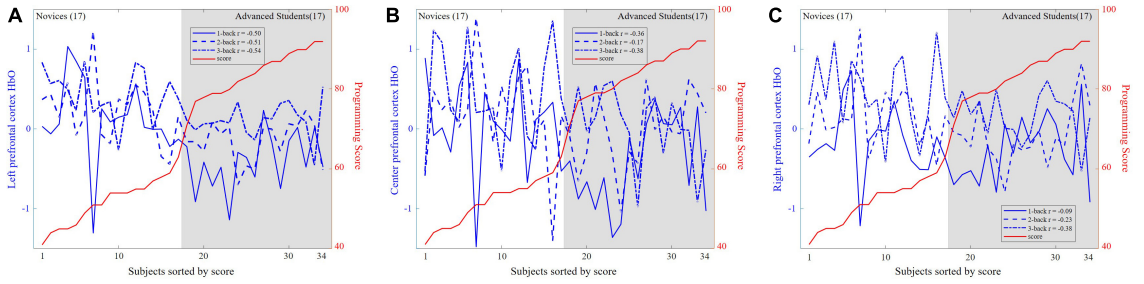


FIGURE 7
Correlations between hemoglobin (HbO) and programming score for the three n-back levels in the left, center, and right prefrontal cortices. **(A)** Left prefrontal cortex. **(B)** Center prefrontal cortex. **(C)** Right prefrontal cortex.

consumed more mental resources. In comparison with the novices, the advanced students illustrated lower prefrontal cortex activation for the n-back task, which was considered to place less of a demand on working-memory load and was easier to complete for the advanced students. This is the underlying basis for the measurement of prefrontal cortex activation of fNIRS signals during n-back tasks to predict an individual's programming ability.

The results also indicated the possibility of predicting students' programming potential through general psychological experiments such as n-back test (as shown in **Figure 5**). This method may be especially useful for evaluating individuals with no programming foundation.

Additionally, HbR activation reduced slightly during the n-back task, as illustrated in **Figure 6**, and programming ability showed no main effect on HbR. This may be because relative to HbO, the HbR concentration is weak and difficult to detect in real time (Abibullaev and An, 2012), making it harder to detect significant effects on HbR activation in task-based fNIRS (Huppert et al., 2006).

Although the main effects of programming ability on HbO were significantly different in the left, middle, and right prefrontal cortices, by applying the deep learning method of attention-based CNN to the fNIRS signals, we found that the channels that can better distinguish programming ability were in the left region of the prefrontal cortex (see **Figure 8**). CNNs can extract microfeatures from temporal domain signals that may be corrupted by statistical analysis. CNNs can also be used to discover and extract the appropriate internal structure through convolution and pooling operations and automatically generate the deep features of the raw data. Moreover, the deep features are robust against translation and scaling (Zhao et al., 2017); they work well in discarding noisy series and can extract meaningful patterns while ignoring patterns without value (Aussem and Murtagh, 1997). By introducing attention modules (i.e., the SE block), we can open the black box to see which feature the CNN network relies on to identify those signals. As the CNN network is gradually fitted, the attention modules indicate the important channels, as shown in **Figure 8**. These high-weight channels are those that the neural network uses to understand and classify. In other words, these channels and their corresponding brain regions have higher resolution in this task. The SE block can also improve the representational power of the regular CNN by offering it a kind of dynamic channel-wise fixing feature (Hu et al., 2020). Furthermore, the feature importance values produced by the self-attention operation can be used for model pruning, which can lead to the construction of more efficient physiological signal analysis networks.

The results of the current study also provided further evidence to support the lateralization of brain functions. The left prefrontal cortex was more important in programming ability prediction, as demonstrated in **Figures 7**, **8**. Many studies have reported functional hemispheric asymmetry in cognitive processes (Gazzaniga, 1989, 1995; Goel, 2019). Smith et al. (1996) and Smith and Jonides (1997) used PET technology to study the neural basis of working memory with the n-back paradigm. Their results showed that the activation areas in the verbal and spatial n-back tasks are different: the former activates the left hemisphere, and the latter activates the right hemisphere. Baddeley and Logie (1999) also reviewed the evidence showing that the left hemisphere is associated with verbal working-memory tasks. As shown in **Figure 6**, our results also demonstrated that left prefrontal cortex regional activation was more dynamic during the verbal n-back test.

This study had some limitations that should be addressed in future studies. First, we did not consider the mediating factors between working memory and programming ability. Previous studies have shown that the relationship between working memory and academic performance is mediated by visuospatial abilities (Logie et al., 2000) and the ability to control attention (Kane and Engle, 2003). Future studies should aim to control these mediating factors to acquire more rigorous results. Second, the number of participants who met the criteria for advanced students was relatively small. Further studies with additional data are required to improve the generalizability of the findings.

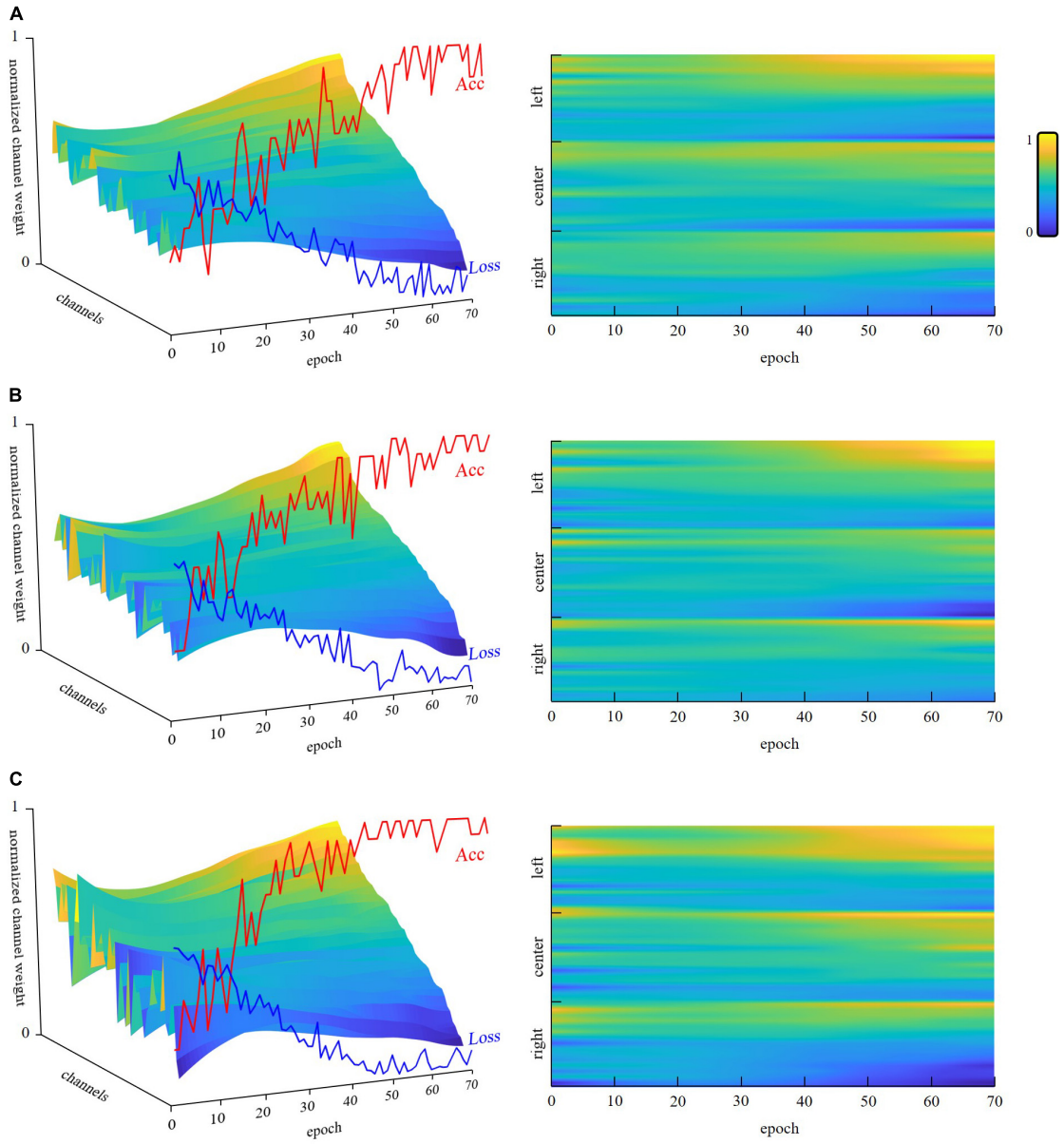


FIGURE 8
The channel weights fitting with the training process of the attention-based convolutional neural network (CNN). The left panel shows all channel weights sorted in the order of the final training results. All channels are sorted by the last loss value. The right panel presents a top-down view of the left panel. However, the channels are re-ranked into the left, center, and right regions. We list the three groups in **Supplementary Table 1**. The right panel has two axes: re-ranked channels and epochs. The normalized channel weight is represented by color bars. **(A)** 1-back, **(B)** 2-back, **(C)** 3-back.

Conclusion

To the best of our knowledge, few empirical studies have directly examined the relationship between working-memory capacity and programming ability, and no studies have provided direct neural evidence to support this relationship. The present study attempts to fill this gap and demonstrates that students' programming ability can be predicted by evaluation of their working-memory capacity while providing direct neural evidence supporting this relationship. The results of our analyses indicate that fNIRS detected functional neural changes associated with the workload in the prefrontal cortex, demonstrating that the hemodynamic responses measured in the prefrontal cortex can be used to discriminate between novices and advanced students. Additionally, we utilized an attention-based CNN to analyze the spatial domains of the fNIRS signals and demonstrated that the left prefrontal cortex

was more important than other brain regions for programming ability prediction.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/GiantMushroom/NIRdataset.git.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. The participants provided their written informed consent to participate in this study.

Author contributions

XG: methodology and writing – original draft preparation. YL: conceptualization, validation, writing – review and editing, and funding acquisition. YZ: formal analysis and writing – original draft preparation. CW: investigation and data curation. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Research Projects of the Humanities and Social Sciences Foundation of the Ministry of Education of China (grant no. 20YJA880034).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.1058609/full#supplementary-material

References

Abibullaev, B., and An, J. (2012). Classification of frontal cortex haemodynamic responses during cognitive tasks using wavelet transforms and machine learning algorithms. *Med. Eng. Phys.* 34, 1394–1410. doi: 10.1016/j.medengphy.2012.01.002

Alloway, T. P. (2009). Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *Eur. J. Psychol. Assess.* 25, 92–98. doi: 10.1027/1015-5759.25.2.92

Anmarkrud, Ø, Andresen, A., and Bråten, I. (2019). Cognitive load and working memory in multimedia learning: conceptual and measurement issues. *Educ. Psychol.* 54, 61–83. doi: 10.1080/00461520.2018.1554484

Asgher, U., Ahmad, R., Naseer, N., Ayaz, Y., Khan, M. J., and Amjad, M. K. (2019). Assessment and classification of mental workload in the prefrontal cortex (PFC) using fixed-value modified beer-lambert law. *IEEE Access* 7, 143250–143262. doi: 10.1109/access.2019.2944965

Aussem, A., and Murtagh, F. (1997). Combining neural network forecasts on wavelet-transformed time series. *Connect. Sci.* 9, 113–122. doi: 10.1080/095400997116766

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Baddeley, A. (2010). Working memory. *Curr. Biol.* 20, R136–R140. doi: 10.1016/j.cub.2009.12.014

Baddeley, A. D., and Hitch, G. (1974). Working memory. *Psychol. Learn. Motiv.* 8, 47–89. doi: 10.1016/s0079-7421(08)60452-1

Baddeley, A. D., and Logie, R. H. (1999). "Working memory: the multiple-component model," in *Models of Working Memory*, eds A. Miyake and P. Shah (Cambridge, MA: Cambridge University Press), 28–61.

Bai, S., Kolter, J. Z., and Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1803.01271 (accessed April 19, 2018).

Barrouillet, P., and Lépine, R. (2005). Working memory and children's use of retrieval to solve addition problems. *J. Exp. Child Psychol.* 91, 183–204. doi: 10.1016/j.jecp.2005.03.002

Borovykh, A., Bohte, S., and Oosterlee, C. W. (2018). Dilated convolutional neural networks for time series forecasting. *J. Comput. Finan.* 22, 73–101. doi: 10.21314/jcf.2019.358

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., and Noll, D. C. (1997). A parametric study of prefrontal cortex involvement. *Neuroimage* 5, 49–62.

Buitrago Flórez, F., Casallas, R., Hernández, M., Reyes, A., Restrepo, S., and Danies, G. (2017). Changing a generation's way of thinking: teaching computational thinking through programming. *Rev. Educ. Res.* 87, 834–860. doi: 10.3102/0034654317710096

Cantin, R. H., Gnaedinger, E. K., Gallaway, K. C., Hesson-McInnis, M. S., and Hund, A. M. (2016). Executive functioning predicts reading, mathematics, and

theory of mind during the elementary years. *J. Exp. Child Psychol.* 146, 66–78. doi: 10.1016/j.jecp.2016.01.014

Chance, B., Zhuang, Z., UnAh, C., Alter, C., and Lipton, L. (1993). Cognition-activated low-frequency modulation of light absorption in human brain.pdf. *Proc. Natl. Acad. Sci. U.S.A.* 90, 3770–3774. doi: 10.1073/pnas.90.8.3770

Chen, Y., Kang, Y., Chen, Y., and Wang, Z. (2019). Probabilistic forecasting with temporal convolutional neural network. *arxiv* [Preprint]. doi: 10.48550/arXiv.1906.04397

Choi, J., Kim, J., Hwang, G., Yang, J., Choi, M., and Bae, H. (2016). Time-divided spread-spectrum code-based 400 fW-detectable multichannel fNIRS IC for portable functional brain imaging. *IEEE J. Solid State Circ.* 51, 484–495. doi: 10.1109/jssc.2015.2504412

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., et al. (1997). Temporal dynamics of brain activation during a working memory task.pdf. *Nature* 386, 604–608. doi: 10.1038/386604a0

Dunst, B., Benedek, M., Jauk, E., Bergner, S., Koschutnig, K., Sommer, M., et al. (2014). Neural efficiency as a function of task demands. *Intelligence* 42, 22–30. doi: 10.1016/j.intell.2013.09.005

Ferrari, M., and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63, 921–935. doi: 10.1016/j.neuroimage.2012.03.049

Franceschini, M. A., Joseph, D. K., Huppert, T. J., Diamond, S. G., and Boas, D. A. (2006). Diffuse optical imaging of the whole head. *J. Biomed. Opt.* 11:054007.

Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *J. Exp. Psychol. Gen.* 141, 2–18. doi: 10.1037/a0024338

Gazzaniga, M. (1989). Organization of human brain. *Science* 245, 947–952.

Gazzaniga, M. (1995). Principles of human brain organization derived from split-brain studies. *Neuron* 14, 217–228. doi: 10.1016/0896-6273(95)90280-5

Genc, E., Fraenz, C., Schluter, C., Friedrich, P., Hossiep, R., Voelkle, M. C., et al. (2018). Diffusion markers of dendritic density and arborization in gray matter predict differences in intelligence. *Nat. Commun.* 9:1905. doi: 10.1038/s41467-018-04268-8

Goel, V. (2019). Hemispheric asymmetry in the prefrontal cortex for complex cognition. *Handb. Clin. Neurol.* 163, 179–196. doi: 10.1016/B978-0-12-804281-6.00010-0

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: a survey. *Comput. Vis. Media* 8, 331–368. doi: 10.1007/s41095-022-0271-y

Haier, R. J., Siegel, B. V., Nuechterlein, K. H., Hazlett, E., Wu, J. C., and Paek, J. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence* 12, 199–217. doi: 10.1016/j.neubiorev.2009.04.001

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2013). Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. *Front. Hum. Neurosci.* 7:935. doi: 10.3389/fnhum.2013.00935

Ho, T. K. K., Gwak, J., Park, C. M., and Song, J.-I. (2019). Discrimination of mental workload levels from multi-channel fNIRS using deep leaning-based approaches. *IEEE Access* 7, 24392–24403. doi: 10.1109/access.2019.2900127

Hocke, L. M., Oni, I. K., Duszynski, C. C., Corrigan, A. V., Frederick, B. D., and Dunn, J. F. (2018). Automated processing of fNIRS data-a visual guide to the pitfalls and consequences. *Algorithms* 11:67. doi: 10.3390/a11050067

Hong, K.-S., and Yaqub, M. A. (2019). Application of functional near-infrared spectroscopy in the healthcare industry: a review. *J. Innov. Opt. Health Sci.* 12:1930012. doi: 10.1142/s179354581930012x

Hoshi, Y., and Tamura, M. (1993). Detection of dynamic changes in cerebral oxygenation coupled to neuronal function mental work in man. *Neurosci. Lett.* 150, 5–8. doi: 10.1016/0304-3940(93)90094-2

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intellig.* 42, 2011–2023.

Huppert, T. J., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2009). HomER: A review of time-series analysis methods for near-infrared spectroscopy of the brain. *Appl. Opt.* 48, 280–298.

Huppert, T. J., Hoge, R. D., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2006). A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *Neuroimage* 29, 368–382. doi: 10.1016/j.neuroimage.2005.08.065

Isbilir, E., Cakir, M. P., Acarturk, C., and Tekerek, A. S. (2019). Towards a multimodal model of cognitive workload through synchronous optical brain imaging and eye tracking measures. *Front. Hum. Neurosci.* 13:375. doi: 10.3389/fnhum.2019.00375

Ivanova, A. A., Srikant, S., Sueoka, Y., Kean, H. H., Dhamala, R., O'Reilly, U. M., et al. (2020). Comprehension of computer code relies primarily on domain-general executive brain regions. *eLife* 9:e58906. doi: 10.7554/eLife.58906

Janani, A., Sasikala, M., Chhabra, H., Shajil, N., and Venkatasubramanian, G. (2020). Investigation of deep convolutional neural network for classification of motor imagery fNIRS signals for BCI applications. *Biomed. Signal Process. Control* 62:102133. doi: 10.1016/j.bspc.2020.102133

Kane, M. J., and Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *J. Exp. Psychol. Gen.* 132, 47–70. doi: 10.1037/0096-3445.132.1.47

Khoe, H. C. H., Low, J. W., Wijerathne, S., Ann, L. S., Salgaonkar, H., Lomanto, D., et al. (2020). Use of prefrontal cortex activity as a measure of learning curve in surgical novices: results of a single blind randomised controlled trial. *Surg. Endosc.* 34, 5604–5615. doi: 10.1007/s00464-019-07331-7

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information.pdf. *J. Exp. Psychol.* 55, 352–358. doi: 10.1037/h0043688

Koh, P. H., Glaser, D. E., Flandin, G., Kiebel, S., Butterworth, B., Maki, A., et al. (2007). Functional optical signal analysis: a software tool for near-infrared spectroscopy data processing incorporating statistical parametric mapping. *J. Biomed. Opt.* 12:064010. doi: 10.1117/1.2804092

Logie, R. H., Della Sala, S., Wynn, V., and Baddeley, A. D. (2000). Visual similarity effects in immediate verbal serial recall. *Q. J. Exp. Psychol. A* 53, 626–646. doi: 10.1080/713755916

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., et al. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Med. Image Anal.* 20, 112–134. doi: 10.1016/j.media.2014.10.011

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., et al. (2014). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Trans. Biomed. Eng.* 62, 1120–1131. doi: 10.1109/TBME.2014.2369495

Maki, A., Yamashita, Y., Ito, Y., Watanabe, E., Mayanagi, Y., and Koizumi, H. (1995). Spatial and temporal analysis of human motor activity using noninvasive NIR topography. *Med. Phys.* 22, 1997–2005. doi: 10.1118/1.597496

Matthes, K., and Gross, F. (1938). Fortlaufende registrierung der lichtabsorption des blutes in zwei verschiedenen spektralbezirken. *Naunyn Schmiedebergs Arch. Pharmacol.* 191, 381–390.

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., and Berman, M. G. (2021). Load-dependent relationships between frontal fNIRS activity and performance: a data-driven PLS approach. *Neuroimage* 230:117795. doi: 10.1016/j.neuroimage.2021.117795

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., and Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. *Neuroimage* 11, 424–446. doi: 10.1006/nimg.2000.0572

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* 25, 46–59. doi: 10.1002/hbm.20131

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., et al. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Ann. N.Y. Acad. Sci.* 1464, 5–29. doi: 10.1111/nyas.13948

Purpura, D. J., and Ganley, C. M. (2014). Working memory and language: skill-specific or domain-general relations to mathematics? *J. Exp. Child Psychol.* 122, 104–121. doi: 10.1016/j.jecp.2013.12.009

Quaresima, V., and Ferrari, M. (2019). A mini-review on functional near-infrared spectroscopy (fnirs): where do we stand, and where should we go? *Photonics* 6:87. doi: 10.3390/photonics6030087

Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., et al. (2002). Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. *Neuropsychology* 16, 370–379. doi: 10.1037/0894-4105.16.3.370

Relkin, E., de Ruiter, L. E., and Bers, M. U. (2021). Learning to code and the acquisition of computational thinking by young children. *Comput. Educ.* 169:104222. doi: 10.1016/j.compedu.2021.104222

Saadati, M., Nelson, J., and Ayaz, H. (2020). "Convolutional neural network for hybrid fNIRS-EEG mental workload classification," in *Advances in Neuroergonomics and Cognitive Engineering*, ed. H. Ayaz (Cham: Springer), 221–232. doi: 10.3389/fnbot.2022.873239

Sassaroli, A., and Fantini, S. (2004). Comment on the modified Beer-Lambert law for scattering media. *Phys. Med. Biol.* 49, N255–N257. doi: 10.1088/0031-9155/49/14/n07

Shute, V. J. (1995). Who is likely to acquire programming skills? *J. Educ. Comput. Res.* 7, 1–24. doi: 10.2190/vqjd-t1yd-5wvb-rypj

Singh, A. K., and Dan, I. (2006). Exploring the false discovery rate in multichannel NIRS. *Neuroimage* 33, 542–549. doi: 10.1016/j.neuroimage.2006.06.047

Smith, E. E., and Jonides, J. (1997). Working memory a view from neuroimaging. *Cogn. Psychol.* 33, 5–42.

Smith, E. E., Jonides, J., and Koeppe, R. A. (1996). Dissociating verbal and spatial working. *Cerebr. Cortex* 6, 11–20. doi: 10.1093/cercor/6.1.11

Strangman, G., Culver, J. P., Thompson, J. H., and Boas, D. A. (2002). A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *Neuroimage* 17, 719–731. doi: 10.1006/nimg.2002.1227

Swanson, H. L., and Alloway, T. P. (2012). "Working memory, learning, and academic achievement," in *APA Educational Psychology Handbook, Vol 1: Theories, Constructs, and Critical Issues*, eds K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, and J. Sweller (Cham: Springer), 327–366.

Tong, Y., Lindsey, K. P., and deB Frederick, B. (2011). Partitioning of physiological noise signals in the brain with concurrent near-infrared spectroscopy and fMRI. *J. Cereb. Blood Flow Metab.* 31, 2352–2362.

Trakoolwilaiwan, T., Behboodi, B., Lee, J., Kim, K., and Choi, J. W. (2017). Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: three-class classification of rest, right-, and left-hand motor execution. *Neurophotonics* 5:011008. doi: 10.1117/1.NPh.5.1.011008

Tu, J.-J., and Johnson, J. R. (1990). Can computer programming improve problem-solving ability? *SIGCSE Bull.* 22, 30–33.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA.

Villringer, A., Planck, J., Hock, C., Schleinkofer, L., and Dimagl, U. (1993). Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neurosci. Lett.* 154, 101–104.

Werner, L., Denner, J., Campe, S., and Kawamoto, D. C. (2012). "The fairly performance assessment: measuring computational thinking in middle school," in *SIGCSE '12 Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, Raleigh.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philos. Trans. A Math. Phys. Eng. Sci.* 366, 3717–3725. doi: 10.1098/rsta.2008.0118

Wolf, M., Wolf, U., Toronov, V., Michalos, A., Paunescu, L. A., Choi, J. H., et al. (2002). Different time evolution of oxyhemoglobin and deoxyhemoglobin concentration changes in the visual and motor cortices during functional stimulation: a near-infrared spectroscopy study. *Neuroimage* 16(3 Pt 1), 704–712. doi: 10.1006/nimg.2002.1128

Yang, D., Hong, K. S., Yoo, S. H., and Kim, C. S. (2019). Evaluation of neural degeneration biomarkers in the prefrontal cortex for early identification of patients with mild cognitive impairment: an fNIRS study. *Front. Hum. Neurosci.* 13:317. doi: 10.3389/fnhum.2019.00317

Yang, D., Huang, R., Yoo, S. H., Shin, M. J., Yoon, J. A., Shin, Y. I., et al. (2020). Detection of mild cognitive impairment using convolutional neural network: temporal-feature maps of functional near-infrared spectroscopy. *Front. Aging Neurosci.* 12:141. doi: 10.3389/fnagi.2020.00141

Yeung, M. K., Lee, T. L., Han, Y. M. Y., and Chan, A. S. (2021). Prefrontal activation and pupil dilation during n-back task performance: a combined fNIRS and pupillometry study. *Neuropsychologia* 159:107954. doi: 10.1016/j.neuropsychologia.2021.107954

Yu, F., and Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. *arxiv* [Preprint]. Available online at: https://arxiv.org/abs/1511.07122 (accessed April 30, 2016).

Zhao, B., Lu, H., Chen, S., Liu, J., and Wu, D. (2017). Convolutional neural networks for time series classification. *J. Syst. Eng. Electron.* 28, 162–169. doi: 10.21629/jsee.2017.01.18

frontiers | Frontiers in Neuroimaging

TYPE Original Research
PUBLISHED 16 February 2023
DOI 10.3389/fnimg.2023.1119539

OPEN ACCESS

EDITED BY
Shu Zhang,
Northwestern Polytechnical University, China

REVIEWED BY
Xiaowei Yu,
University of Texas at Arlington, United States
Sergio Luiz Novi Junior,
Western University, Canada

*CORRESPONDENCE
J. Jean Chen
jchen@research.baycrest.org

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

RECEIVED 08 December 2022
ACCEPTED 20 January 2023
PUBLISHED 16 February 2023

CITATION
Agrawal V, Zhong XZ and Chen JJ (2023) Generating dynamic carbon-dioxide traces from respiration-belt recordings: Feasibility using neural networks and application in functional magnetic resonance imaging. *Front. Neuroimaging* 2:1119539. doi: 10.3389/fnimg.2023.1119539

Generating dynamic carbon-dioxide traces from respiration-belt recordings: Feasibility using neural networks and application in functional magnetic resonance imaging

Vismay Agrawal[1], Xiaole Z. Zhong[1,2] and J. Jean Chen[1,2,3]*

[1]Baycrest Centre for Geriatric Care, Rotman Research Institute, Toronto, ON, Canada, [2]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada, [3]Department of Biomedical Engineering, University of Toronto, Toronto, ON, Canada

Introduction: In the context of functional magnetic resonance imaging (fMRI), carbon dioxide (CO_2) is a well-known vasodilator that has been widely used to monitor and interrogate vascular physiology. Moreover, spontaneous fluctuations in end-tidal carbon dioxide ($PETCO_2$) reflects changes in arterial CO_2 and has been demonstrated as the largest physiological noise source for denoising the low-frequency range of the resting-state fMRI (rs-fMRI) signal. However, the majority of rs-fMRI studies do not involve CO_2 recordings, and most often only heart rate and respiration are recorded. While the intrinsic link between these latter metrics and CO_2 led to suggested possible analytical models, they have not been widely applied.

Methods: In this proof-of-concept study, we propose a deep-learning (DL) approach to reconstruct CO2 and PETCO2 data from respiration waveforms in the resting state.

Results: We demonstrate that the one-to-one mapping between respiration and CO_2 recordings can be well predicted using fully convolutional networks (FCNs), achieving a Pearson correlation coefficient (r) of 0.946 ± 0.056 with the ground truth CO_2. Moreover, dynamic $PETCO_2$ can be successfully derived from the predicted CO_2, achieving r of 0.512 ± 0.269 with the ground truth. Importantly, the FCN-based methods outperform previously proposed analytical methods. In addition, we provide guidelines for quality assurance of respiration recordings for the purposes of CO_2 prediction.

Discussion: Our results demonstrate that dynamic CO_2 can be obtained from respiration-volume using neural networks, complementing the still few reports in DL of physiological fMRI signals, and paving the way for further research in DL based bio-signal processing.

KEYWORDS

deep learning, fully convoluted neural network, carbon dioxide, respiratory variability, functional MRI, physiological signal analysis, cerebrovascular reactivity (CVR)

1. Introduction

Carbon dioxide (CO_2) is a potent vasodilator used that has been shown to rely mainly on the nitric oxide pathway to increase arterial diameter (Pelligrino et al., 1999; Najarian et al., 2000; Peebles et al., 2008; Iadecola, 2017). Blood-vessel diameter is highly sensitive to the surrounding CO_2 concentration, with increasing CO_2 partial pressures leading to linear increases in both vessel diameter and flow (Hülsmann and Dubelaar, 1988; Komori et al., 2007). In Komori et al. for example, this increase was shown to be 21.6% for arteriolar diameter and 34.5% flow velocity for a 50% change in CO_2 partial pressure in rabbit arterioles (Komori et al., 2007). The partial pressure of carbon dioxide (PCO_2) is the measure of CO_2 within arterial or venous blood. It often

serves as a marker of sufficient alveolar ventilation within the lungs. Under normal physiologic conditions, the value of PCO_2 ranges between 35 and 45 mmHg, or 4.7–6.0 kPa. Typically the measurement of PCO_2 is performed *via* arterial blood gas, but the end-tidal pressure of CO_2 ($PETCO_2$) is related to intravascular PCO_2 through a linear relationship under steady-state conditions (Peebles et al., 2007, 2008), allowing arterial PCO_2 to be estimated from $PETCO_2$.

Dynamic CO_2 recordings have multiple utilities and implications. In the past decades, the CO_2-driven functional magnetic resonance imaging (fMRI) response has been the preeminent method for mapping cerebrovascular reactivity (Blockley et al., 2017; Chen, 2018; Chen and Gauthier, 2021). Wise et al. first reported the contribution of spontaneous fluctuations in arterial PCO_2 to the resting-state fMRI (Wise et al., 2004). Chang et al. followed up this work by demonstrating the potential relationship between $PETCO_2$ and respiratory-volume variability (RVT) (Chang and Glover, 2009). Using recordings of spontaneous $PETCO_2$ variations, Golestani et al. determined the fMRI response function that links $PETCO_2$ to the resting-state blood-oxygenation level dependent (BOLD) signal (Golestani et al., 2015), and also demonstrated $PETCO_2$ as the primary source of physiological noise in resting-state BOLD. It has even been used to demonstrate the possible existence of neuronally-motivated vascular networks in the brain (Bright et al., 2020). Furthermore, Chan et al. (2021) found that PCO_2 (not $PETCO_2$) fluctuations also contribute significantly to resting-state BOLD signal variability (Chan et al., 2020). While the mid-breath PCO_2 does not reflect intravascular PCO_2, $PETCO_2$ does provide a quantitative estimate of arterial PCO_2, and is more widely used in fMRI experiments for the purposes of denoising (Murphy et al., 2013) and CVR mapping (Pinto et al., 2020). The substantial influence of dynamic $PETCO_2$ fluctuations on resting-state (Golestani and Chen, 2020) and dynamic functional connectivity has been demonstrated recently (Nikolaou et al., 2016). Dynamic CO_2 can also allow vascular lag structures to be estimated, providing an important metric for assessing vascular health (Champagne et al., 2019). Given the unique variance explained by PCO_2 and $PETCO_2$, it is safe to say that dynamic CO_2 is a useful thus desirable metric for those working with resting-state fMRI data.

Despite the increasing realization of the value of CO_2 recordings, it is often impossible to obtain recordings of CO_2 during an fMRI session. Most study sites are not equipped with an MRI-compatible capnometer that also facilitates continuous recording of PCO_2. Moreover, the many thousands of legacy fMRI data sets (e.g., Human Connectome Project, UK Biobank) certainly do not include CO_2 recordings. On the other hand, respiratory volume variations, which had previously been related to $PETCO_2$ variations, are more readily available thanks to the incorporation of respiratory-volume belts in modern MRI systems. RVT was first introduced by Birn et al. as a noise source in fMRI that introduces unique signal variability (Birn et al., 2006). Today, while RVT measurements during fMRI sessions are increasingly common, they are still unavailable in large-scale studies and legacy data sets. As a possible solution, recent work by Salas et al. (2020) demonstrated that the RVT time series can in principle be reconstructed from fMRI data using a convolutional neural network (CNN).

Chang et al. previously showed that $PETCO_2$ can be related to RVT through a respiratory-response function (Birn et al., 2008). However, this relationship has been difficult to reproduce in resting-state conditions, as we will show with our data. In the resting state, not only is it impossible to derive quantitative CO_2 values from respiratory volume, it is also difficult to obtain a deterministic relationship between dynamic patterns of respiratory volume and CO_2 variation. Thus, in this study, we also use the principle of DL, but our focus is to bridge the gap between respiratory and CO_2 recordings. Our aim is to demonstrate the feasibility of using DL to produce dynamic CO_2 waveforms from the respiratory time series.

1.1. Background on neural networks

In the majority of DL methods for neuroimaging, 2D inputs are used to produce 2D outputs (Zhu et al., 2019). Image-to-image translation is used for cross-modality conversion, denoising, super-resolution and reconstruction (Kaji and Kida, 2019). Our problem entails the estimation of a 1D signal from another 1D signal, and within this context, past research has used convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Traditional CNNs consist of convolutional layers followed by fully connected layers (dense layers) terminating the network (Rawat and Wang, 2017). As CNNs are the most successful type of DL model for 2D image analysis, and physiological signals are 1D time-series data, some have converted 1D signals to 2D data to be fed into a CNN, and have obtained good results (Shah et al., 2022). The advantage of using 1D CNNs over 2D CNNs and RNNs is the significant reduction in the number of training parameters, which is helpful when the training data is limited (as the application at hand). Applications of 1D CNNs include ECG classification and anomaly detection in biomedical signals (Kiranyaz et al., 2021). Salas et al. pioneered the use of 1D CNN for estimating physiological fluctuations in fMRI, an application closely related to ours. They segmented the BOLD fMRI signals into fixed time-windows and fed them into a CNN, where the dense layer predicted a single point of the respiration waveform at the center of the window. To predict the entire time series, all the time-windows have to be separately propagated through the network, entailing high complexity and computational cost. Moreover, commonly found respiration-belt recordings have variable lengths, which are incompatible with the use of dense layers.

In this work, we implemented a type of CNN known as fully convolutional networks (FCNs) (Long et al., 2015). A FCN is simply a traditional CNN without any fully connected layers. Fully convolutional layers in FCN permit the use of variable-length input and also minimizes the computational cost. Previously, a 1D U-net (a type of FCN that includes skip connections) was implemented for reconstructing low-frequency respiratory-volume signals from fMRI time-series data (Bayrak et al., 2020). Here, we demonstrate the use of simple FCNs (without skip connections) for predicting 1D data wherein the encoder-decoder architecture exploits the latent space to streamline the prediction of CO_2 traces from respiration-belt signal, in the presence of limited training data.

2. Methods

2.1. Data acquisition

We recorded percent-CO_2 (%CO2) fluctuations and respiratory bellows simultaneously in a group of 18 healthy adults (age 20–38 years) using the Biopac System (Biopac Inc., Goleta, CA, USA). The Biopac respiration belt was positioned below the ribcage, and detects

respiratory depth by sensing abdominal circumference changes. $\%CO_2$ data were acquired through gas lines attached to masks affixed to subjects' faces. The Biopac $\%CO_2$ module (CO2100C) is calibrated to measure $\%CO_2$ concentration in the range of 0 to 10%. In total, the available data set consisted of 136 resting-state recordings from different subjects, which were 10.8 min long on average (min = 7.2 min, max = 16.1 min). The procedure was approved by the Research Ethics Board of Baycrest (REB# 11–47, approved Dec. 2011–19). To the best of our knowledge, this is the largest data set of its kind in existence.

2.2. Data preprocessing

The preprocessing steps consist of (1) low-pass filtering both respiration and CO_2 waveforms (f < 1 Hz) and (2) correcting the delay between $\%CO_2$ and respiration signal by cross-correlation. The low pass filter's cutoff frequency was determined based on the respiratory rate of an individual (0.2–0.4 Hz). The delay between $\%CO_2$ and respiration waveforms were corrected by shifting the $\%CO_2$ time course by the time lag yielding the maximum negative cross-correlations between it and the respiration waveform. We found that across all cases, to achieve this, the %CO2 time course had to be shifted to the left (backwards in time) by an average of 8.5 s (with a standard deviation of 1.5 s).

After the delay correction process, we rejected data that yielded absolute Pearson correlations of <0.4. Recordings were also rejected if their length was <3 min, too short to allow adequate training. More details on the correlation and data-length threshold are given in the quality assurance section. The respiration belt data was in arbitrary units; hence it was normalized by subtracting the temporal mean and dividing the result by standard deviation. The same procedure was applied to the $\%CO_2$ waveforms. Further details about the normalization are provided in the next subsection. Both the waveforms were then resampled to 10 Hz and exported in CSV format to be later imported during the training phase of the neural network.

To obtain $PETCO_2$ from the normalized $\%CO_2$ recordings, the peak-detection step [available through SciPy: (Virtanen et al., 2020)] ensures the minimum distance between the two peaks is twice the sampling interval. In other words, we assumed the time between two exhales is at least 2 s, which is consistent with our recorded respiratory intervals (3–5 s per breath). Moreover, the lower limit of the amplitude of the peak was set to be 0.3, and negative peaks are also rejected.

2.2.1. Data normalization

As previously mentioned, both $\%CO_2$ and respiration-belt data were demeaned and normalized to unit standard deviation (such that SD = 1). The respiration data is fluctuations in voltage transduced from expansions and contractions of the belt. As such, it varies with slight variations in belt tightness and positioning, and needs to be normalized across subjects to achieve inter-subject consistency. In part due to the need of using normalized respiration as the independent variable, this latter would encode no quantitative $\%CO_2$ information. That is, there could be a many-to-one relationship between normalized respiration and unnormalized CO_2. To mitigate this issue, we demeaned and normalized the $\%CO_2$ time series in the same manner. In this manuscript, all the further mentions of CO_2 denote normalized $\%CO_2$, unless stated otherwise.

2.2.2. Quality assurance

A critical part of successful application of machine learning is quality assurance (QA) of the training and testing data. It is more probable to find noise in respiration data, wherein artifacts such as subject movement and talking can easily confound respiration-belt recordings. Moreover, if the participant does not consistently breathe from the abdomen, the respiration belt data may not correspond well with the CO_2 data. During the data-collection phase, useful precautions include ensuring that the respiration belt and CO_2 gas lines are properly connected. Such precautions not only reduce the unwanted waveforms but also increase the feasibility of machine-learning approaches. To discard the undesirable recordings, we have evaluated our data based on the criteria below. Nonetheless, it is informative to use data containing some level of noise and artifact for the purposes of representativeness. Therefore, the threshold used in the rejection process is generously selected.

2.2.2.1. Length of the recording

In general, for our approach, longer data sets are more desirable. It was observed that all the recordings were either <3 min or more than 6 min in length, drawing a clear distinction between test recordings and usable recordings. Thus, the lower limit for the time length was set to 3 min. Figure 1 shows the histogram plot of all the recordings after the time-length thresholding.

2.2.2.2. Pearson correlation coefficient

As previously mentioned, Pearson's correlation (r) between the respiration belt and CO_2 time courses is used for initial QA purposes. The threshold for the absolute value of correlation between CO_2 and respiration is −0.4, as respiratory volume and CO_2 are expected to be negatively associated. This limit was empirically determined through manual review of the recordings. Figure 2 shows that even though the threshold was −0.4, there were no recordings with r between −0.4 and −0.5, only one recording with r = −0.5 and most of the recordings had an r value of <-0.6.

2.2.2.3. Low-frequency noise in the waveforms

Within the 0.1–0.5 Hz frequency band, noise in the respiratory and CO_2 waveforms can impair our ability to relate the two waveforms, even if the recording-duration and correlation-coefficient thresholds are met. Such noise most likely originates from faulty attachment of the respiration belt and from drifts in the recording modules. As it could potentially overlap with breathing frequency, it cannot be separated from the signal by using filters. However, this type of noise can be identified through a mismatch in the low-frequency portion (<0. 2Hz) of the power spectra of CO_2 and respiration, as shown in Supplementary Figure 1. This type of noise is also reflected in the signal time series as periodic decreases or increases in the amplitude of signal. Conversely, an exemplary data set is shown in Supplementary Figure 2.

2.2.3. Neural network

Obtaining the CO_2 concentration from the respiration waveform is a 1D-to-1D (time series to time series) translation problem, which is modeled using a 1D fully convolutional encoder-decoder

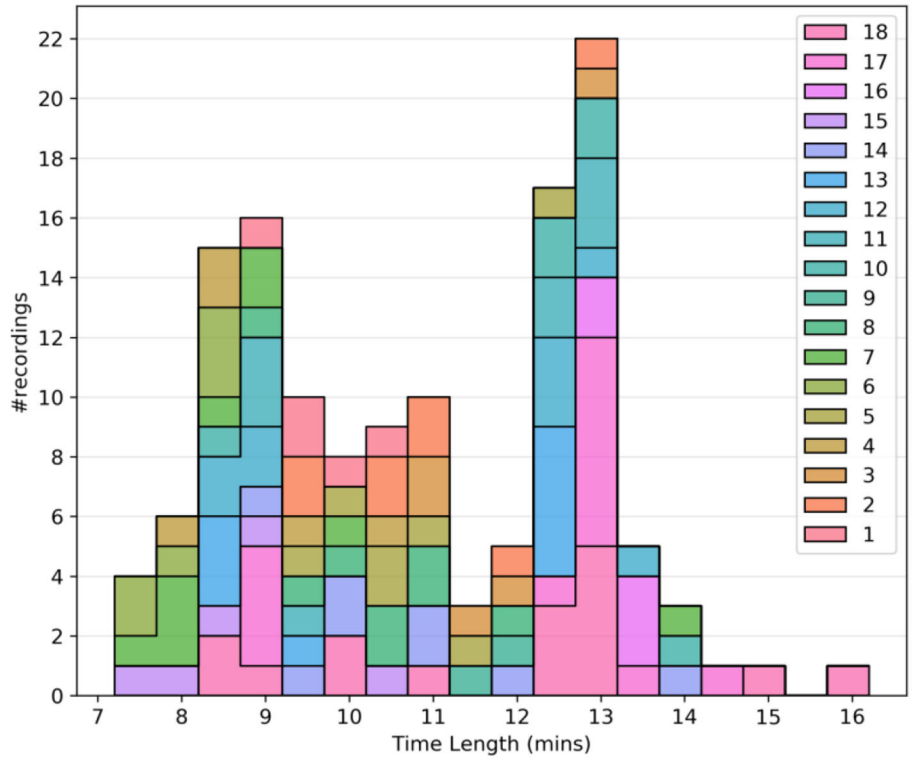


FIGURE 1
Quality assurance metrics: Histogram plot of the time length of recordings after time length thresholding. Different colors are used to separate the subjects.

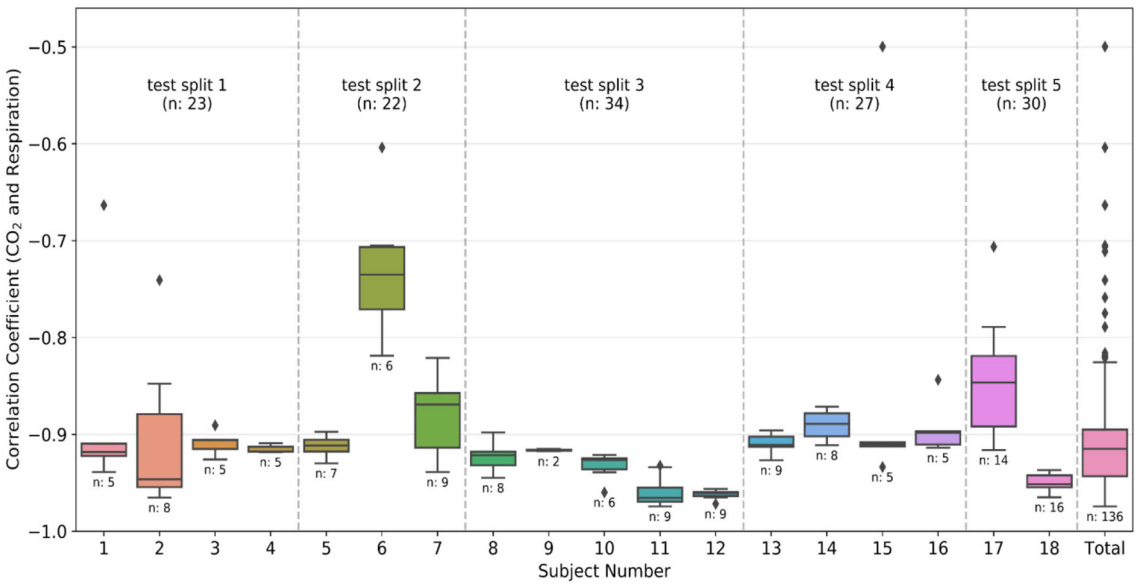


FIGURE 2
Quality assurance metrics: Box plots of the correlation coefficient between CO_2 and respiration waveforms from each individual subject and the total data after preprocessing. The number of recordings available for each subject is also given below the box plot. The divisions created by the dashed line show the groups made during the k-fold split of the dataset. The group number is the same as the test split number, and the total number of recordings in the group is also provided in the plot. The color-coding is the same as Figure 1.

architecture. This modeling is analogous to prevalent image-to-image translation or semantic segmentation using 2D FCNs (Long et al., 2015; Alotaibi, 2020). However, most recent works in image-to-image translation problems involve adversarial training (Pang et al., 2022), which is notoriously hard especially with limited data. Thus, adversarial training is excluded in this paper.

Constructing a deep neural network often involves trial and error for tuning hidden layers. To find an optimum number of hidden layers in the network, several FCNs architectures are investigated, until overfitting was observed (test phase error increases with increasing network complexity). All codes are written in Python and use the PyTorch library, and would be publicly available on GitHub.

2.2.3.1. FCN architecture

Input to the network was an array of size C x L, where the number of input channels, C = 1 and L is the length of recording. Although the respiration recordings were normalized using standard deviation, the resultant data range still varied between data sets. To bound the respiration amplitude within a fixed range, the respiration array was further normalized using the tanh operator before being passed on to the fully convolutional layers. We implemented four different FCN architectures, each having one (FCN-1L), two (FCN-2L), four (FCN-4L) and six (FCN-6L) convolution layers, respectively, between the input and output layers.

FCN-1L consists of a single convolution operation with a kernel of length 7 and replicate padding of 3 on both sides (head and tail) of the input waveform. The kernel length is chosen to balance model complexity with prediction accuracy. FCN-2L encodes the tanh-normalized respiration waveform by convolving it with a 4×7 kernel (4 kernels of length 7) with a stride of 2, which means the input is downsampled by a factor of 2. This is followed by ReLU nonlinearity (activation function) and finally a transposed convolution to decode the hidden layer into CO_2. Both the convolution and transposed convolution are performed with a stride of 2, which replaces the need for a pooling layer to downsample the output of convolutional layers and an unpooling layer to upsample the output of transposed convolutional layers. Similarly, FCN-4L consists of 2 convolution and 2 transposed convolutional layers, and FCN-6L architecture adds another 1 layer to both encoder and decoder sections. The network architecture of FCN-4L is shown in Figure 3.

2.2.3.2. Loss function

We also experimented with two different loss functions. The first loss function is the mean squared error (MSE) computed between the measured and predicted CO_2 waveforms, which is widely used in regression problems (Equation 1). However, as the regression was performed between the waveforms of pseudo-periodic nature, it was observed that the network learned to predict zero-crossings extremely well, but the extremities were left underfitted, lowering the scores of $PETCO_2$ predictions. To rectify this problem, a second loss function, the weighted MSE (MSE_{Wgt}), was introduced Equation 2), with the weights set to the normalized amplitudes of the ground truth CO_2 waveform for each timepoint. The weighting provides higher preference to the peaks, and hence we hypothesized that it would provide better results for $PETCO_2$.

$$MSE = \frac{1}{L}\sum_{i=1}^{L}(y_i - \hat{y}_i)^2 \quad (1)$$

$$MSE_{Wgt} = \frac{1}{L}\sum_{i=1}^{L}[(y_i - \hat{y}_i)/|y_i|]^2 \quad (2)$$

where, y_i and y_i are the predicted and ground truth CO_2 respectively for the i^{th} time point, and L is the length of the recording. Networks trained with the weighted cost function are denoted by the postfix "-Wgt."

2.2.4. Training

The 18 subjects were split into 5 subsets (splits), and the training was executed using the k-fold cross-validation strategy. It is typical to use either 10-fold or 5-fold cross-validation as it generally results in a model with low bias, modest variance and low-computational cost compared to leave-one-out cross-validation strategy (Rodriguez et al., 2010). In our dataset, as the number of subjects is relatively limited, we opted for k = 5, and each time one subset was left out from the training phase to be used in testing the accuracy of the network. Each subject can have multiple recordings, and the data was divided based on the subjects (and not recordings) to ensure that the training and testing data has no scans sharing a common subject. The divisions created by dotted lines in Figure 2 correspond to the different splits. As visible in the figure, the splits contain data from 2, 5, 4, 4, and 3 subjects, yielding total numbers of 30, 34, 27, 23, and 22 recordings, respectively. Each split has a different number of total recordings, which enhances the generalizability of the results. We implemented two training strategies.

2.2.4.1. Method 1. Equal-length data segments

In this method, we formatted the training data as an array of equal-sized data segments obtained by segmenting the input recordings. As the training was performed on a GPU, the computation parallelized in the tensor with multiple batches, reducing the training time. We used the chunk size of 90 s and a batch size of 256. The drawback of this method is the unavoidable error introduced due to edge effects during convolution, which is proportional to the number of chunks.

2.2.4.2. Method 2. Variable-length data segments

In this method the input array length could be of variable sizes. The drawback of using variable-length input is that it prevents us from grouping the data in batches for parallel processing in the GPU. On the positive note, unlike in Method 1, Method 2 precludes the segmenting-induced edge effects. We implemented both methods. The training time was <20 s irrespective of the network type or training method. All the networks were trained using Adam optimizer for 15 epochs. Hyperparameters corresponding to the optimizer like learning rate and decay rate were fine-tuned manually for each network. In total, we trained four FCNs, each using two loss functions, on the 5-fold split data. The training was performed on a 12GB GeForce GTX TITAN X GPU. All networks used <500MB GPU memory during the training phase.

2.2.4.3. Reference methods

To the best of our knowledge, there have been no previous attempts to derive the CO_2 waveform from respiratory traces using machine learning. To establish the performance of our approach against a possible alternative, we employed two reference methods. First, based on previous work by Chang and Glover (2009), defining a $PETCO_2$ as the convolution of RVT with RRF (and then normalized, negated and shifted temporally for maximum cross-correlation). This is referred to as the RVTRRF method, described by Equation 3. RVT was estimated from respiration waveform as detailed in Birn et al. (2008).

$$PETCO_2'(t) = RVTRRF(t) = RVT(t)^{*}RRT(t) \quad (3)$$

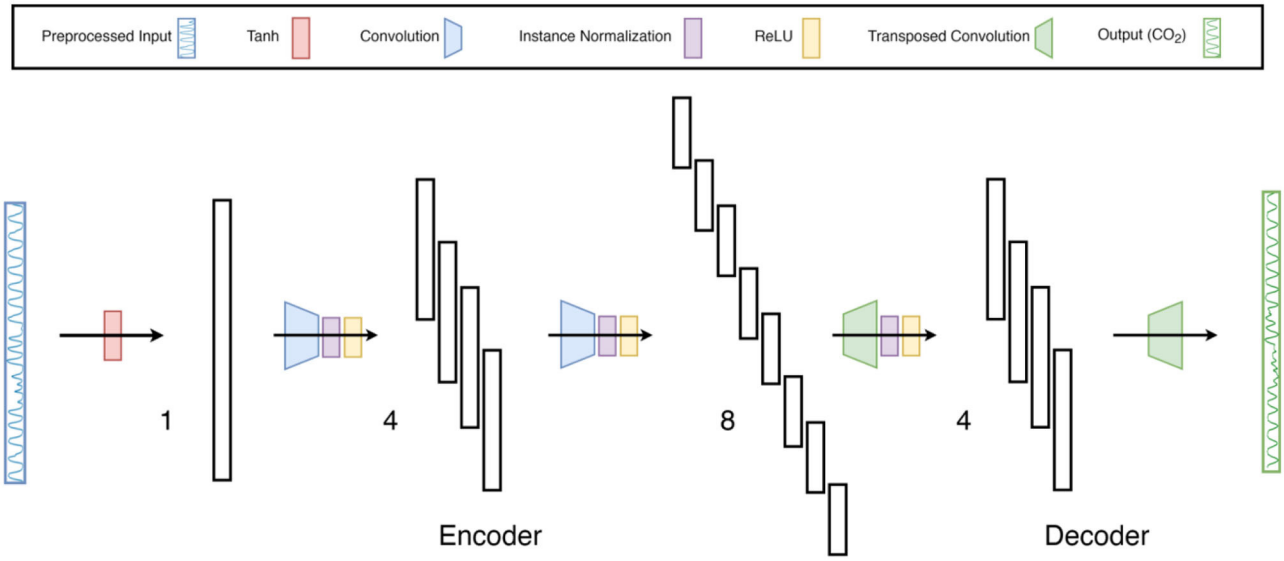


FIGURE 3
Neural-network architecture: 4-Layer Fully Convolutional Network. The architecture shown here is a type of encoder-decoder neural network consisting of fully convolutional layers, followed by instance normalization and ReLu non-linearity. The last layer does not contain normalization and activation function as it is a regression problem. Moreover, the input is first normalized using tanh activation function to constrain the input data between −1 and 1. The numbers 1, 4 and 8 indicate the number of filters per layer.

where $PETCO_2'(t)$ is the estimated $PETCO_2$. RRF is the respiratory response function, and * denotes convolution. Similar to what was done previously (Chang and Glover, 2009), at the testing stage, we corrected the lag between RVTRRF [$PETCO_2^{'}(t)$] and $PETCO_2$ using the maximum cross-correlation between the two signals, where the time shift was allowed to vary between −120 and 120 s. Moreover, to maintain the scaling of $PETCO_2$ as obtained from neural networks, we normalized and demeaned RVTRRF with the standard deviation and mean of $PETCO_2$.

Second, defining a linear-regression (LR) model relating CO_2 to respiratory volume (Equation 4), and $PETCO_2$'(t) is extracted from the CO_2 time courses (measured using the Biopac system in this case).

$$CO_2'(t) = \beta \cdot Resp(t) + \varepsilon \tag{4}$$

where CO_2' is the estimated CO_2, Resp(t) is the respiratory-belt signal, ε is the intercept, and β is the linear weighting factor derived from the "training data," and the LR model could be understood as a single convolutional operation with a unit kernel size, making it similar to a machine learning linear regression problem. The training and testing partitioning are as described for the FCNs. MSE loss function was backpropagated similar to the FCNs.

2.2.4.4. Evaluation criteria

For the evaluation, the Pearson correlation coefficient (r), mean squared error (MSE), mean absolute error (MAE) (Equation 5) and mean absolute percent error (MAPE) (Equation 6) were calculated between (1) predicted CO_2 and ground-truth CO_2, (2) predicted $PETCO_2$ and ground-truth $PETCO_2$. As the MAPE is sensitive to zero crossings, it was only calculated between the predicted $PETCO_2$ and ground-truth $PETCO_2$.

$$MAE = \frac{1}{L}\sum_{i=1}^{L}\left(|y_i - \hat{y}_i|\right) \tag{5}$$

$$MAPE = \frac{1}{L}\sum_{i=1}^{L}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \tag{6}$$

We also performed statistical comparisons amongst correlation coefficients and MSE values obtained using all FCN and reference methods using the Kruskal-Wallis test, corrected for false-discovery rate.

The final validation is inspired by a practical application of CO_2 recordings, namely examining the relationship between $PETCO_2$ and resting-state fMRI time series. For this we include 3 cases acquired from each of the 2 healthy young subjects (male, age = 25 and 33 years). All data were acquired using a Siemens TIM Trio 3 T system and a 32-channel head coil. CO_2 was acquired during these scans as described earlier. That is, each dataset contains the following:

- Case 1: spin-echo EPI, TR = 323 ms, TE = 45 ms, flip angle = 90°, 2,082 frames, voxel size = X: 3.48 mm, Y: 3.48 mm, Z: 6.25 mm;
- Case 2: gradient-echo EPI, TR = 323 ms, TE = 30 ms, 2,230 frames, voxel size = X: 3.48 mm, Y: 3.48 mm, Z: 6.25 mm;
- Case 3: simultaneous multi-slice gradient-echo EPI, TR = 323 ms, TE = 30 ms, flip angle = 40°, 2,230 frames, voxel size = X: 3.48 mm, Y: 3.48 mm, Z: 6 mm;

Preprocessing steps include: (1) filtering to 0.01–0.1 Hz band with AFNI (Cox, 1996); (2) spatial smoothing with a 5 mm kernel (Jenkinson et al., 2012) (3) Discard the first 5 volumes in each scan to allow the brain to reach a steady state. All recorded and FCN-generated CO_2 and $PETCO_2$ time courses were low-pass filtered

to 0.01–0.1 Hz to match the temporal resolution of the respective fMRI data.

3. Results

Results for two representative data sets are shown in Figure 4. Method 1 (equal data length) adds no extra benefit to the training process and results in poor performance due to possible truncation effects in training data. Thus, all the results provided here correspond to Method 2. The results are shown in Figure 4 and summarized in Table 1. The best method, as determined by the lowest error terms (MSE, MAE, MAPE) and highest Pearson correlation (r) is indicated in bold. The predicted and ground-truth $PETCO_2$ show excellent visual agreement for FCN-4L-Wgt (Figure 4B). From Table 1, we can see that the CO_2 estimation error obtained from FCN-4L and FCN-4L-Wgt architecture are identical, with the errors corresponding to $PETCO_2$ being slightly lower in the latter case. Since r is unaffected by scaling and translation, and since the LR model involves only scaling and translation, the modeling step would not improve r. Strangely, the RVTRRF model performs worse than the LR model (for $PETCO_2$), suggesting that estimating $PETCO_2$ from the peaks of the CO_2 (and hence respiration) waveform may be more robust.

Figure 5 shows the r distribution across the entire test dataset for one of the five splits. The LR method is outperformed by all FCN methods (and significantly so by FCN-4L-Wgt) for CO_2 prediction. The difference between FCN-4L and FCN-4L-Wgt is not noticeable in the case of CO_2 prediction, but overall, FCN-4L-Wgt achieved the highest r values, while FCN-6L achieved the lowest r variability. However, for $PETCO_2$, FCN-4L-Wgt reached higher r values than did FCN-4L, demonstrating the superiority of a weighted loss function. FCN-6L performs worse than all the other FCN networks for $PETCO_2$ prediction. However, these differences are not statistically significant, as can also be seen in Table 2, in which every approach is compared to the apparent leader (FCN-4L-Wgt). Note that the RVTRRF method only reached a maximum r score of just below 0.5, substantially lower compared to all FCN networks. As previously mentioned, the r scores for RVTRRF correspond to maximum cross correlation with $PETCO_2$, thus the scores are always positive. There is no such limitation for the FCNs, resulting in some network correlation coefficients in the distribution.

Figure 6 compares the correlation scores between training and testing phase for all the networks. From these plots, it can be inferred that FCN-6L likely overfits the training data, as reflected by a worse performance than that of the other networks (as reflected by a lower r). Since FCN-4L performs better than FCL-2L and doesn't show huge differences between training and testing results, we can say four convolutional blocks are the optimum number for our given training data. Moreover, in our best model, MAPE score for $PETCO_2$ is 0.142 (< 0.2), reflective of good prediction performance.

Figure 7 compares the correlation coefficients across the five splits for all the networks. The r-score ranking in the case of CO_2 prediction does not match with that of $PETCO_2$ prediction. In the case of CO_2, the r for FCN-4L-Wgt closely resemble those of FCN-4L, but the former performed better for $PETCO_2$ (in all but one split). Though the best model varied depending on the split number and varies between CO_2 and $PETCO_2$ prediction, FCN-4L-Wgt consistently outperformed other models, exemplified in part by the highest correlation coefficients. The inter-split variability in r is the lowest for the reference methods (RVTRRF and LR) and highest for FCN methods, the various FCN methods themselves do not appear to exhibit different degrees of inter-split performance variability. Moreover, the performance rankings of the various methods are consistent across the splits and in line with the trends observed in Figure 5. Combining the results of Figure 7 with the information in Figure 2, it can be seen that the poor CO_2-prediction performance for all methods across the second split is due to one subject (subject 6). CO_2 prediction in Split 3 was best overall. Yet, the LR model performs worst in predicting $PETCO_2$ in the 3rd split, reflecting that higher correlation between CO_2 and respiration does not necessarily translate into higher correlation between $PETCO_2$ and respiration. This point is further demonstrated by contrasting r scores of $PETCO_2$ and CO_2 for the LR approach in the remaining splits.

Figure 8 demonstrates the application of the FCN-4L-predicted dynamic $PETCO_2$, which have established correlation with the resting-state fMRI signal. We show that the $PETCO_2$-fMRI correlation maps for the ground-truth and predicted $PETCO_2$ are highly similar in all scan sessions (Cases 1, 2 and 3) and subjects (Datasets 1 and 2). This preliminary demonstration suggests promise in using the model-predicted $PETCO_2$ for fMRI applications.

4. Discussion

As a proof-of-concept study, we demonstrated that it is feasible to use an FCN to predict dynamic CO_2 from respiration variations. Furthermore, the performance of the FCN surpasses that of regression and convolution-based methods. Note that the results only pertain to dynamic patterns in CO_2, not to absolute CO_2, which cannot be predicted from non-quantitative respiration traces alone. Nonetheless, possible applications range from improving the feasibility of breath-holding based fMRI studies (Murphy et al., 2013) that lack CO_2 recordings, to the use of the CO_2-O_2 exchange ratio for vascular reactivity mapping (Chan et al., 2020). These applications do not require quantitative values of CO_2 and $PETCO_2$.

4.1. Machine learning in physiological signal processing

The use of machine learning and DL models is prevalent in physiological signal data such as electromyogram (EMG), electroencephalogram (EEG), electrocardiogram (ECG), and electrooculogram (EOG) (Rim et al., 2020). It has been continuously observed that DL models perform better than other, classical machine learning models. Rim et al. conducted a review of 147 studies using DL in EMG, ECG, EEG, EOG and their combinations (Rim et al., 2020), and concluded that most were in the domain of classification, feature-extraction and data compression, wherein CNN, RNN, CNN+RNN models were most commonly used. The studies were divided into 3 categories. The first category exploits machine-learning models to extract features followed by DNN as a classifier to boost the accuracy of classification by obtaining useful features from raw data. The second involves DL as a feature extractor and traditional machine learning as a classifier to reduce hand-crafted labeling of the dataset. The third strategy uses an end-to-end DL pipeline to train raw data and receive the final output to build a robust model for the above-mentioned tasks. Due to

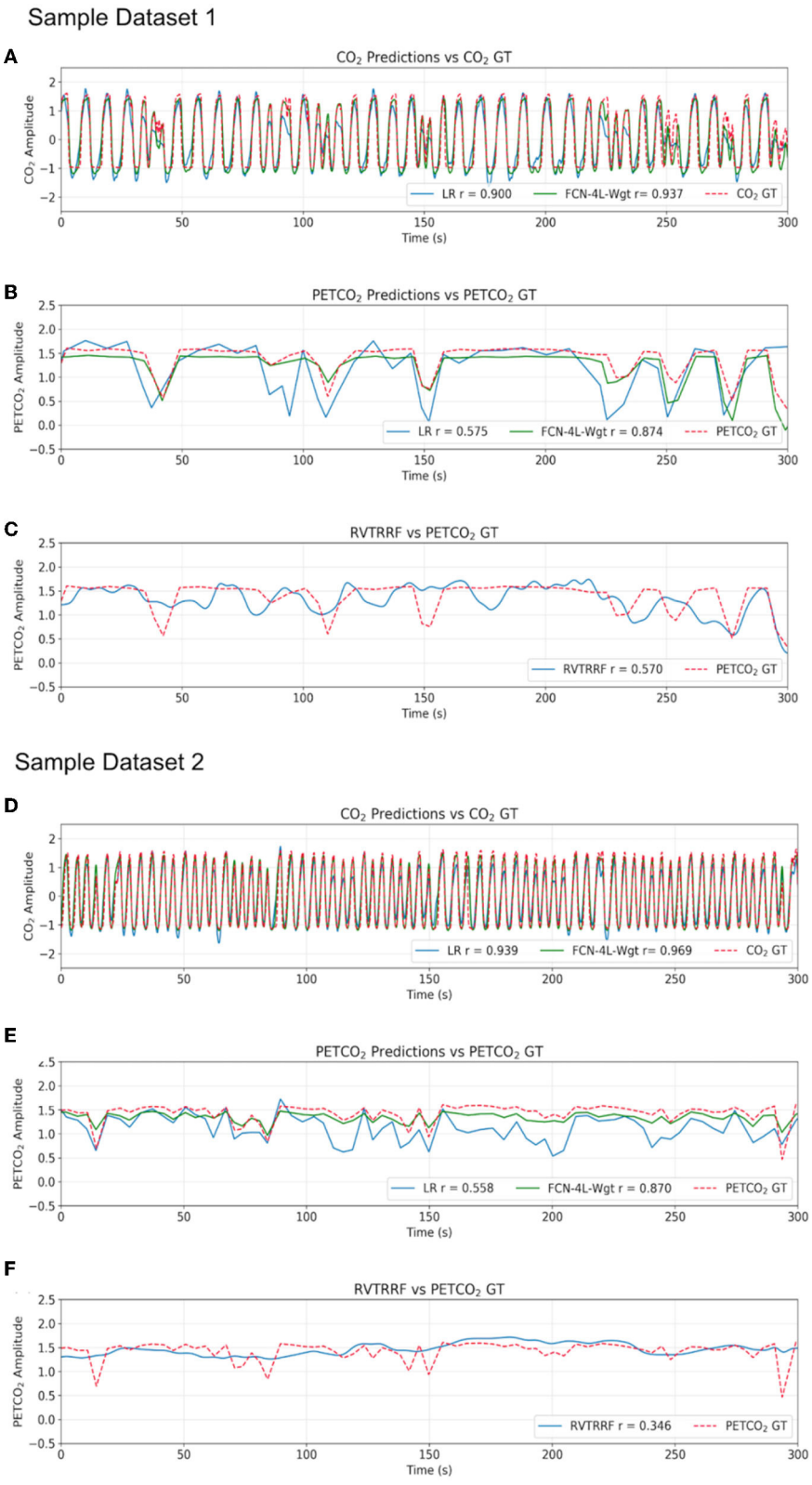


FIGURE 4
Qualitative comparison of resultant outputs. Two different sample predictions are shown from the test dataset, and for each of the example, comparisons are made between **(A, D)** the CO_2 prediction and ground truth (GT), **(B, E)** the $PETCO_2$ prediction from the reference linear regression model (LR), FCN-4L-Wgt model and the GT, and **(C, F)** $PETCO_2$ estimated from RVTRRF and the $PETCO_2$ GT.

TABLE 1 Quantitative assessment of various approaches and network structures.

Average across all 5 splits	RVTRRF	LR	FCN-1L	FCN-2L	FCN-4L	FCN-6L	FCN-4L-Wgt
r CO_2	-	0.901 ± 0.061	0.931 ± 0.055	0.922 ± 0.06	0.946 ± 0.054	0.944 ± 0.055	0.946 ± 0.056
r $PETCO_2$	0.256 ± 0.132	0.311 ± 0.239	0.443 ± 0.261	0.45 ± 0.262	0.5 ± 0.266	0.461 ± 0.235	0.512 ± 0.269
MSE CO_2	-	0.19 ± 0.103	0.138 ± 0.094	0.151 ± 0.101	0.108 ± 0.097	0.11 ± 0.097	0.106 ± 0.101
MSE $PETCO_2$	0.032 ± 0.028	0.026 ± 0.021	0.02 ± 0.018	0.019 ± 0.017	0.018 ± 0.017	0.02 ± 0.018	0.017 ± 0.017
MAE CO_2	-	0.337 ± 0.079	0.269 ± 0.077	0.276 ± 0.08	0.223 ± 0.076	0.227 ± 0.081	0.213 ± 0.08
MAE $PETCO_2$	0.121 ± 0.055	0.112 ± 0.045	0.094 ± 0.04	0.093 ± 0.039	0.081 ± 0.035	0.085 ± 0.038	0.08 ± 0.036
MAPE $PETCO_2$	0.125 ± 0.109	0.112 ± 0.084	0.094 ± 0.077	0.095 ± 0.078	0.085 ± 0.073	0.089 ± 0.074	0.084 ± 0.077

RVTRRF, RVT convolved with RRF; LR, linear regression; FCN-XL, "X" layered FCN used; -Wgt, with weighted MSE cost function. The parameters used in the assessment include: the correlation coefficient (r), the mean-squared error (MSE), the mean absolute error (MAE) and the mean-absolute percent error (MAPE). Each metric was calculated for every recording in the test set across all five splits. The mean and standard deviation (mean ± std) were calculated for all the metrics in each test split. Likewise, the average of (mean ± std) was taken across all the 5 splits and displayed in this table.

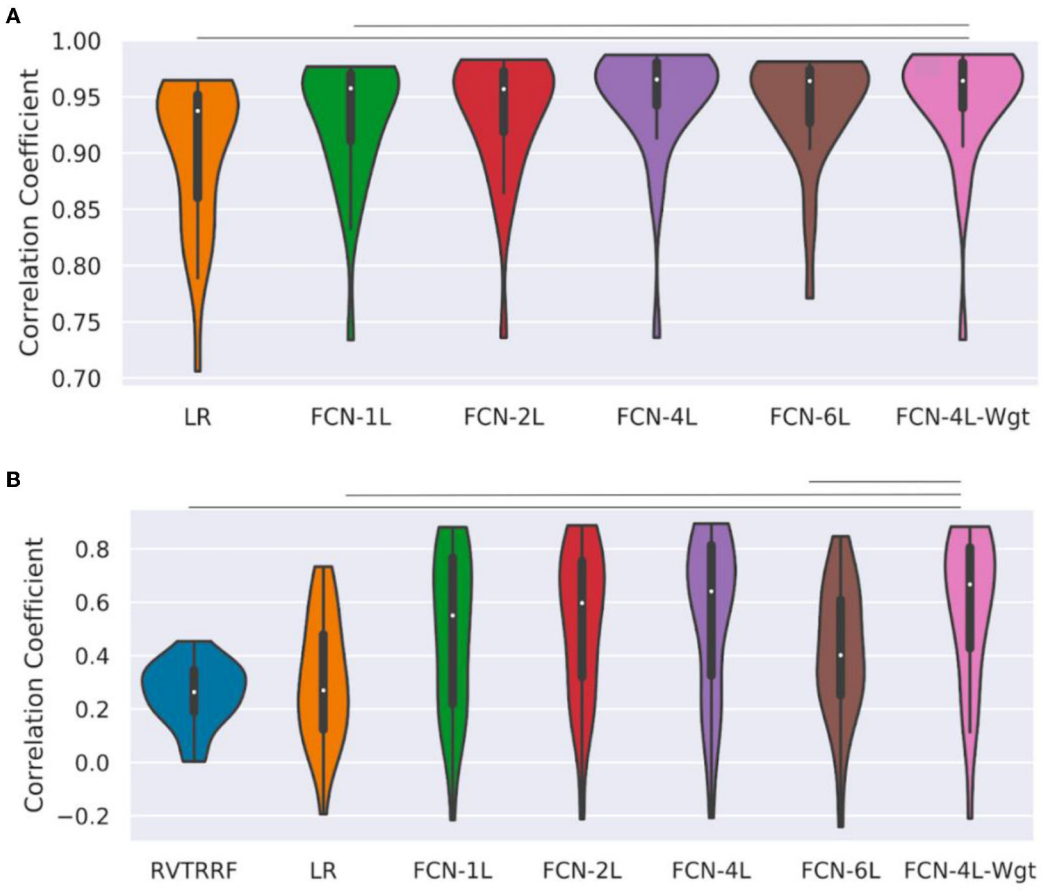


FIGURE 5
Performance of different methods: Distribution of correlation coefficients (r) on test dataset, where r is computed between **(A)** ground-truth and predicted CO_2, and **(B)** the ground-truth and predicted $PETCO_2$ obtained on the test dataset (for one of the five splits) is compared for different models used in the study and shown in the form of a bean plot. The median r for each method is shown as a white dot at the centers of the distributions. The horizontal lines indicate statistically significant differences between the two approaches at the ends of the lines. The FCN-4L-Wgt approach is significantly superior than the RVTRRF and LRF approaches for predicting CO_2, and better than FCN-6L additionally in predicting $PETCO_2$, shown by the significantly higher r values.

the absence of a comparative study involving all 3 methods (Rim et al., 2020), we could not assess the best strategy. Our pipeline is positioned between the second and third categories, as we used an end-to-end DNN to estimate CO_2 as an intermediate step, followed by a post-processing step to obtain the final $PETCO_2$ waveform.

4.2. Utility and current status of using RVT for generating $PETCO_2$

As RVTRRF is correlated with $PETCO_2$, there is a potential of training a convolutional neural network between RVT and

TABLE 2 Statistical comparison of various approaches and network structures with FCN-4L-Wgt.

Metric	RVTRRF	LR	FCN-1L	FCN-2L	FCN-4L	FCN-6L
r CO_2	**0.0001**	0.0287	0.0895	−0.9146	0.2189	0.0001
r $PETCO_2$	**<0.0001**	0.0001	0.1646	0.3593	0.8941	0.0071
MSE CO_2	**<0.0001**	0.0001	0.1646	0.3593	0.8941	0.0071
MSE $PETCO_2$	0.6048	<0.0001	**0.0001**	**0.0005**	0.1833	**0.001**

Listed at the p values indicate the significance of differences. All tests for $PETCO_2$ prediction were performed with 209 degrees of freedom (DOF), and with a DOF of 179 for CO_2 prediction. All p-values were corrected for multiple comparisons. Entries meeting statistical significance are indicated in bold face.

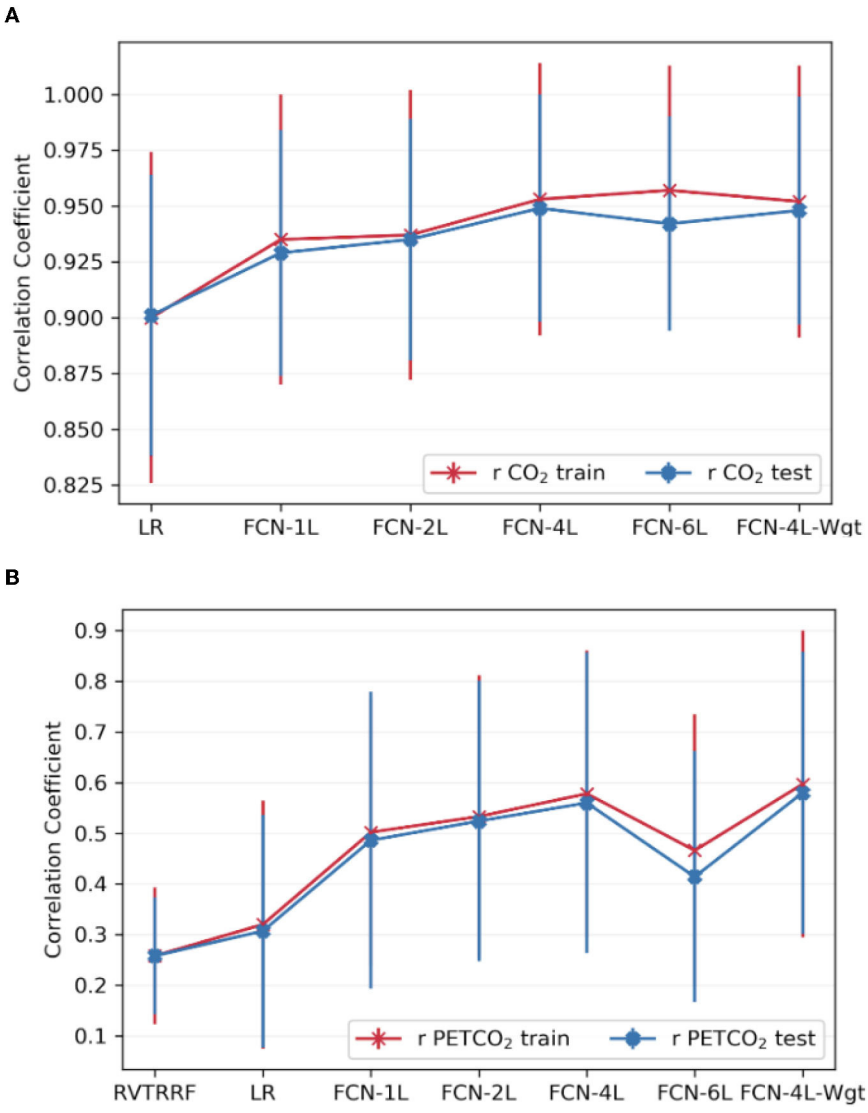


FIGURE 6
Comparison of model performance on train vs. test datasets. The average Pearson correlation coefficient obtained across one of the splits for **(A)** CO_2 and **(B)** $PETCO_2$ between test and train dataset is shown in the top row. The error bars indicate the standard deviation.

$PETCO_2$, which might perform better than a single convolution operation using RRF. This approach aims to find a neural network architecture which could replace the need of RRF. We experimented with different types of neural networks trained to predict $PETCO_2$ from RVT, but none performed adequately. Therefore, we concluded that it is more feasible to design a neural network to associate respiration and CO_2, and predict $PETCO_2$ from CO_2. This may be due to the fact that the latter exploits the evident breathing pattern between respiration patterns and CO_2 and performs well even with limited recording lengths. Conversely, in the former approach, the temporal resolution of RVT is fundamentally constrained to the observed breath durations, and the peak detection algorithm can often miss deep breaths (Power et al., 2020).

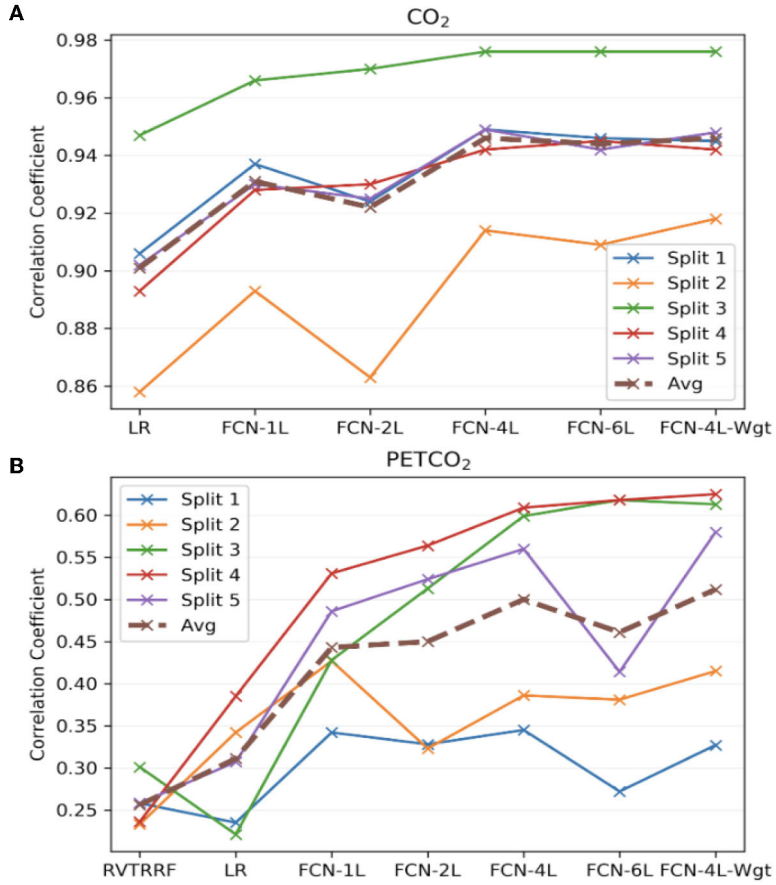


FIGURE 7
Model performance across the five splits. The correlation coefficients (r) obtained across the five splits and their average for all the models, for **(A)** CO_2 and **(B)** $PETCO_2$ prediction. The split number is the same as the splits shown in Figure 2.

As a potential alternative metric of respiratory variability, the windowed respiratory variance (RV), computed as the standard deviation of the respiratory signal over sliding windows of 6 s (Chang et al., 2009), is more robust against noise than RVT as it excludes the influence of breath-cycle duration term. This may however render RV less physiologically related to CO_2. Moreover, the RRF for RV has not been determined (Birn et al., 2008), leading us to exclude the use of RV in this proof-of-principle study. Another potential influence on CO_2 prediction may be the presence of hardware/software filters on the raw recordings. The Biopac system provided software filters to exclude MRI noise (periodicity < 100 ms) while preserving higher physiological frequencies, and it is conceivable that in cases where such frequencies are inadvertently removed from the raw respiratory traces, the ability to predict CO_2 fluctuations may be disadvantaged.

4.3. Other DL architectures

As mentioned previously, a 1D U-net with skip connections had previously been used for translating fMRI data to respiratory-volume data [30]. Skip connections as used in the U-net could be implemented in this study, but as the study is more focused on establishing proof of concept, such complications were avoided in our implementation of FCNs.

There are recently developed alternative network architectures that may also suit our problem. For instance, unpaired and paired image-to-image translation has been accomplished by generative adversarial networks (GANs) such as Pix2Pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017). The translation task is analogous to the task of transforming the respiration-belt data to the CO_2 waveform is analogous. A simple GAN consists of two sub-models, a generator to obtain synthetic samples, and a discriminator to predict the value of the provided sample. The discriminator network in GANs is similar to the explicit loss function used in traditional DL models. In our case, adversarial training would mean that instead of using MSE or weighted MSE loss functions to determine the best CO_2 prediction, another network would distinguish between them. Given that our use case is much simpler, this approach might not add value while incurring higher computational costs and overfitting.

Another alternative are RNNs, such as the long-short term memory (LSTM) (Greff et al., 2017) and gated recurrent unit (GRU) (Zhao et al., 2016) networks, which are widely used in signal processing. At first glance, RNNs seemed a natural choice, but unfortunately, performance was poor (data not shown) for the LSTM. In our implementation, the initial 5-s respiration-signal segment was fed into the LSTM block which would predict the corresponding segment of CO_2 and the hidden state. These outputs along with the next 5-s segment of respiration data were used as the inputs for the next iteration, with the intention that irregularities in breathing would be stored in the network's memory and would help in prediction. Moreover, the 5-s length was comparable to the duration of one breath. Unfortunately, due to the short input-lengths coupled

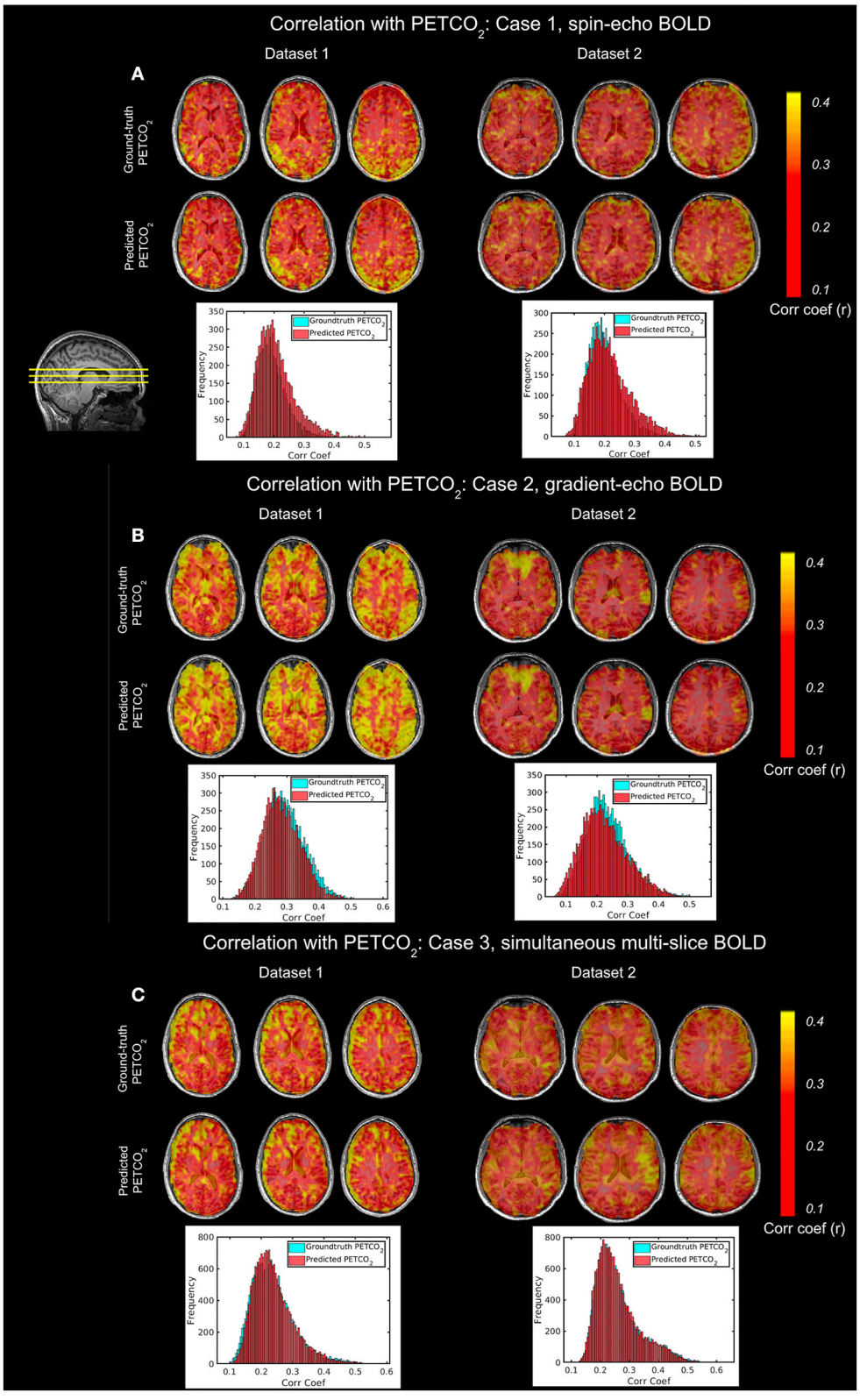


FIGURE 8
Comparison of ground-truth and predicted $PETCO_2$ correlations. Data from 2 different subjects, imaged over multiple sessions [**(A–C)**, respectively] are shown. In each case, the peak cross-correlation maps generated using the ground-truth and predicted $PETCO_2$ time courses are shown in upper and lower rows, with the corresponding correlation-coefficient histograms showing the comparability of the maps. The slice positions are shown by the yellow lines on the sagittal image in the upper-left corner.

with the limited durations of respiration recordings, the concatenated output lacked the smooth transitions between consecutive chunks (i.e. edge effects were apparent in each 5-s block, similar to observed in training method 1), which are required for accurately predicting a slow-varying signal like $PETCO_2$. Thus, we concluded that time-series to time-series translation using RNNs was not feasible unless much longer respiratory and CO_2 recordings were available.

4.4. Limitations

Data quality can be a chief limitation in our approach, and we recommend careful quality assurance as indicated in this work. Another potential limitation is the way in which the test and training data are determined by splitting the full data set; the use of k-fold cross-validation reduces such bias. Peak detection accuracy, which determine the quality of the source $PETCO_2$ data, also needs careful quality assurance. Finally, our method does not attach quantitative values to the estimated PCO_2 or $PETCO_2$ (e.g., in units of mmHg). This is because the quantitative value of $PETCO_2$ depends not only on respiratory patterns, but also on minute ventilation, tidal volume, fitness level, baseline CO_2 storage, and so on (Rawat et al., 2021). Nonetheless, our breath-by-breath CO_2 time series reflects patterns of change are sufficient for fMRI applications.

5. Conclusions

This study demonstrates the feasibility of predicting dynamic $PETCO_2$ from respiration-belt recordings, thus, enabling broader incorporation of $PETCO_2$ in rs-fMRI analysis. Following the successful application of 2D FCNs to image-to-image translation, we introduced 1D FCNs for 1D signal-to-signal translation. The FCN outperformed the analytic regression and convolution models. The study also evaluates the effect of FCN depth as well as the choice of loss function. A 4-layer FCN with weighted MSE performed best across all splits. The results across different deep neural network architectures serve as a literature for further research in signal processing and for the DL community.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, in accordance with institutional guidelines.

Ethics statement

The studies involving human participants were reviewed and approved by Baycrest Research Ethics Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

VA: the conception or design of the work, analysis, interpretation of data for the work, drafting the work or revising it critically for important intellectual content, and final approval of the version to be published. XZ: analysis, interpretation of data for the work, drafting the work and in revising it for important intellectual content, and final approval of the version to be published. JC: the conception or design of the work, interpretation of data for the work, drafting and revising it critically for important intellectual content, and final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This work was supported by the Canadian Institutes of Health Research (CIHR, FDN 148398) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Acknowledgments

We thank Catie Chang of Vanderbilt University (Nashville, USA) for helpful comments.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnimg.2023.1119539/full#supplementary-material

References

Alotaibi, A. (2020). Deep generative adversarial networks for image-to-image translation: a review. *Symmetry* 12, 1705. doi: 10.3390/sym12101705

Bayrak, R. G., Salas, J. A., Huo, Y., and Chang, C. (2020). "A deep pattern recognition approach for inferring respiratory volume fluctuations from fMRI data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Springer International Publishing). 428–436.

Birn, R. M., Diamond, J. B., Smith, M. A., and Bandettini, P. A. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. *Neuroimage* 31, 1536–1548. doi: 10.1016/j.neuroimage.2006.02.048

Birn, R. M., Smith, M. A., Jones, T. B., and Bandettini, P. A. (2008). The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. *Neuroimage* 40, 644–654. doi: 10.1016/j.neuroimage.2007.11.059

Blockley, N. P., Harkin, J. W., and Bulte, D. P. (2017). Rapid cerebrovascular reactivity mapping: enabling vascular reactivity information to be routinely acquired. *Neuroimage* 214–223. doi: 10.1016/j.neuroimage.2017.07.048

Bright, M. G., Whittaker, J. R., Driver, I. D., and Murphy, K. (2020). Vascular physiology drives functional brain networks. *Neuroimage* 116907. doi: 10.1016/j.neuroimage.2020.116907

Champagne, A. A., Bhogal, A. A., Coverdale, N. S., Mark, C. I., and Cook, D. J. (2019). A novel perspective to calibrate temporal delays in cerebrovascular reactivity using hypercapnic and hyperoxic respiratory challenges. *NeuroImage* 187, 154–165. doi: 10.1016/j.neuroimage.2017.11.044

Chan, S. T., Evans, K. C., Song, T. Y., Selb, J., van der Kouwe, A., Rosen, B. R., et al. (2020). Cerebrovascular reactivity assessment with O2-CO2 exchange ratio under brief breath hold challenge. *PLoS ONE* 15, e0225915. doi: 10.1371/journal.pone.0225915

Chan, S. T., Ordway, C., Calvanio, R. J., Buonanno, F. S., Rosen, B. R., and Kwong, K. K. (2021). *Cerebrovascular Responses to O2-CO2 Exchange Ratio Under Brief Breath-Hold Challenge in Patients With Chronic Mild Traumatic Brain Injury*.

Chang, C., Cunningham, J. P., and Glover, G. H. (2009). Influence of heart rate on the BOLD signal: the cardiac response function. *Neuroimage* 44, 857–869. doi: 10.1016/j.neuroimage.2008.09.029

Chang, C., and Glover, G. H. (2009). Relationship between respiration, end-tidal CO_2, and BOLD signals in resting-state fMRI. *Neuroimage* 47, 1381–1393. doi: 10.1016/j.neuroimage.2009.04.048

Chen, J. J. (2018). Cerebrovascular-reactivity mapping using MRI: considerations for Alzheimer's disease. *Front. Aging Neurosci.* doi: 10.3389/fnagi.2018.00170

Chen, J. J., and Gauthier, C. J. (2021). The role of cerebrovascular-reactivity mapping in functional MRI: calibrated fMRI and resting-state fMRI. *Front. Physiol.* 12, 657362. doi: 10.3389/fphys.2021.657362

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

Golestani, A. M., Chang, C., Kwinta, J. B., Khatamian, Y. B., and Chen, J. J. (2015). Mapping the end-tidal CO2 response function in the resting-state BOLD fMRI signal: spatial specificity, test–retest reliability and effect of fMRI sampling rate. *Neuroimage* 104, 266–277. doi: 10.1016/j.neuroimage.2014.10.031

Golestani, A. M., and Chen, J. J. (2020). Controlling for the effect of arterial-CO_2 fluctuations in resting-state fMRI: comparing end-tidal CO_2 clamping and retroactive CO_2 correction. *Neuroimage* 216, 116874. doi: 10.1016/j.neuroimage.2020.116874

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: a search space odyssey. *IEEE Trans. Neural Netw. Learn Syst.* 28, 2222–2232. doi: 10.1109/TNNLS.2016.2582924

Hülsmann, W. C., and Dubelaar, M. L. (1988). Aspects of fatty acid metabolism in vascular endothelial cells. *Biochimie* 70, 681–686. doi: 10.1016/0300-9084(88)90253-2

Iadecola, C. (2017). The neurovascular unit coming of age: a journey through neurovascular coupling in health and disease. *Neuron* 96, 17–42. doi: 10.1016/j.neuron.2017.07.030

Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Kaji, S., and Kida, S. (2019). Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiol. Phys. Technol.* 12, 235–248. doi: 10.1007/s12194-019-00520-y

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1D convolutional neural networks and applications: a survey. *Mech. Syst. Signal Proces.* 151, 107398. doi: 10.1016/j.ymssp.2020.107398

Komori, M., Takada, K., Tomizawa, Y., Nishiyama, K., Kawamata, M., and Ozaki, M. (2007). Permissive range of hypercapnia for improved peripheral microcirculation and cardiac output in rabbits. *Crit. Care Med.* 35, 2171–2175. doi: 10.1097/01.ccm.0000281445.77223.31

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Murphy, K., Birn, R. M., and Bandettini, P. A. (2013). Resting-state fMRI confounds and cleanup. *Neuroimage* 80, 349–359. doi: 10.1016/j.neuroimage.2013.04.001

Najarian, T., Marrache, A. M., Dumont, I., Hardy, P., Beauchamp, M. H., Hou, X., et al. (2000). Prolonged hypercapnia-evoked cerebral hyperemia via K(+) channel- and prostaglandin E(2)-dependent endothelial nitric oxide synthase induction. *Circ. Res.* 87, 1149–1156. doi: 10.1161/01.RES.87.12.1149

Nikolaou, F., Orphanidou, C., Papakyriakou, P., Murphy, K., Wise, R. G., and Mitsis, G. D. (2016). Spontaneous physiological variability modulates dynamic functional connectivity in resting-state functional magnetic resonance imaging. *Philos. Trans. A Math. Phys. Eng. Sci.* 374, 20150183. doi: 10.1098/rsta.2015.0183

Pang, Y., Lin, J., Qin, T., and Chen, Z. (2022). Image-to-image translation: methods and applications. *IEEE Trans. Multimed.* 24, 3859–3881. doi: 10.1109/TMM.2021.3109419

Peebles, K., Celi, L., McGrattan, K., Murrell, C., Thomas, K., and Ainslie, P. N. (2007). Human cerebrovascular and ventilatory CO2 reactivity to end-tidal, arterial and internal jugular vein PCO2. *J. Physiol.* 584, 347–357. doi: 10.1113/jphysiol.2007.137075

Peebles, K. C., Richards, A. M., Celi, L., McGrattan, K., Murrell, C. J., and Ainslie, P. N. (2008). Human cerebral arteriovenous vasoactive exchange during alterations in arterial blood gases. *J. Appl. Physiol.* 105, 1060–1068. doi: 10.1152/japplphysiol.90613.2008

Pelligrino, D. A., Santizo, R. A., and Wang, Q. (1999). Miconazole represses CO(2)-induced pial arteriolar dilation only under selected circumstances. *Am. J. Physiol.* 277, H1484–H1490. doi: 10.1152/ajpheart.1999.277.4.H1484

Pinto, J., Bright, M. G., Bulte, D. P., and Figueiredo, P. (2020). Cerebrovascular reactivity mapping without gas challenges: a methodological guide. *Front. Physiol.* 11, 608475. doi: 10.3389/fphys.2020.608475

Power, J. D., Lynch, C. J., Dubin, M. J., Silver, B. M., Martin, A., and Jones, R. M. (2020). Characteristics of respiratory measures in young adults scanned at rest, including systematic changes and "missed" deep breaths. *Neuroimage* 204, 116234. doi: 10.1016/j.neuroimage.2019.116234

Rawat, D., Modi, P., and Sharma, S. (2021). "Hypercapnea," in *StatPearls.* Treasure Island (FL): StatPearls Publishing.

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

Rim, B., Sung, N.-J., Min, S., and Hong, M. (2020). Deep learning in physiological signal data: a survey. *Sensors* 20, 969. doi: 10.3390/s20040969

Rodriguez, J. D., Perez, A., and Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 569–575. doi: 10.1109/TPAMI.2009.187

Salas, J. A., Bayrak, R. G., Huo, Y., and Chang, C. (2020). Reconstruction of respiratory variation signals from fMRI data. *NeuroImage* 225, 117459. doi: 10.1016/j.neuroimage.2020.117459

Shah, D., Gopan K., G., and Sinha, N. (2022). An investigation of the multi-dimensional (1D vs. 2D vs. 3D) analyses of EEG signals using traditional methods and deep learning-based methods. *Front. Sig. Proc.* 2. 936790. doi: 10.3389/frsip.2022.936790

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Wise, R. G., Ide, K., Poulin, M. J., and Tracey, I. (2004). Resting fluctuations in arterial carbon dioxide induce significant low frequency variations in BOLD signal. *Neuroimage* 21, 1652–1664. doi: 10.1016/j.neuroimage.2003.11.025

Zhao, Y., Li, J., Xu, S., and Xu, B. (2016). "Investigating gated recurrent neural networks for acoustic modeling," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*.

Zhu, G., Jiang, B., Tong, L., Xie, Y., Zaharchuk, G., and Wintermark, M. (2019). Applications of deep learning to neuro-imaging techniques. *Front. Neurol.* 10, 869. doi: 10.3389/fneur.2019.00869

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)*.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 March 2023
DOI 10.3389/fnins.2023.1125666

OPEN ACCESS

EDITED BY
Dajiang Zhu,
University of Texas at Arlington, United States

REVIEWED BY
Zhiguo Luo,
National University of Defense Technology, China
Yimin Hou,
Northeast Electric Power University, China

*CORRESPONDENCE
Fang Hu
✉ cindyljj@163.com
Xieping Gao
✉ xpgao@xtu.edu.cn

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

RECEIVED 16 December 2022
ACCEPTED 22 February 2023
PUBLISHED 09 March 2023

CITATION
Cao C, Li Y, Zhang L, Hu F and Gao X (2023) Identification for the cortical 3-Hinges folding pattern based on cortical morphological and structural features.
Front. Neurosci. 17:1125666.
doi: 10.3389/fnins.2023.1125666

Identification for the cortical 3-Hinges folding pattern based on cortical morphological and structural features

Chunhong Cao[1], Yongquan Li[1], Lele Zhang[1], Fang Hu[2]* and Xieping Gao[3]*

[1]The MOE Key Laboratory of Intelligent Computing and Information Processing, Xiangtan University, Xiangtan, China, [2]Key Laboratory of Medical Imaging and Artificial Intelligence of Hunan Province, Xiangnan University, Chenzhou, China, [3]Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, China

The Cortical 3-Hinges Folding Pattern (i.e., 3-Hinges) is one of the brain's hallmarks, and it is of great reference for predicting human intelligence, diagnosing eurological diseases and understanding the brain functional structure differences among gender. Given the significant morphological variability among individuals, it is challenging to identify 3-Hinges, but current 3-Hinges researches are mainly based on the computationally expensive Gyral-net method. To address this challenge, this paper aims to develop a deep network model to realize the fast identification of 3-Hinges based on cortical morphological and structural features. The main work includes: (1) The morphological and structural features of the cerebral cortex are extracted to relieve the imbalance between the number of 3-Hinges and each brain image's voxels; (2) The feature vector is constructed with the K nearest neighbor algorithm from the extracted scattered features of the morphological and structural features to alleviate over-fitting in training; (3) The squeeze excitation module combined with the deep U-shaped network structure is used to learn the correlation of the channels among the feature vectors; (4) The functional structure roles that 3-Hinges plays between adolescent males and females are discussed in this work. The experimental results on both adolescent and adult MRI datasets show that the proposed model achieves better performance in terms of time consumption. Moreover, this paper reveals that cortical sulcus information plays a critical role in the procedure of identification, and the cortical thickness, cortical surface area, and volume characteristics can supplement valuable information for 3-Hinges identification to some extent. Furthermore, there are significant structural differences on 3-Hinges among adolescent gender.

KEYWORDS

cortical 3-Hinges folding pattern, cortical morphology and structure, gender differences, deep learning, SE-Unet

1. Introduction

Cortical folding patterns quantify the human cerebral cortex, which is highly curled and folded into convex gyri and concave sulci during brain development. From these patterns, we can infer critical clues about cytoarchitecture (Van Essen, 1997; Fischl et al., 2008), neurodevelopment (Dubois et al., 2008), brain function and cognition (Thompson et al., 2004; Jiang et al., 2021). However, because the shapes of the gyri and the sulci are complex

and variable across subjects, it is challenging to quantitatively analyze the cortical folding patterns, estimate precise cross-subject correspondences for them, and establish a mapping from them to brain function and cognition (Fischl et al., 2008). In particular, the location identification of the cortical folding has important clinical reference value for the prediction of human intelligence, the understanding of the brain functional structure (Jiang et al., 2018; Zhang et al., 2018a), and the diagnosis of neurological diseases (Huang et al., 2019).

Despite such difficulty, promising results have been achieved in solving these challenging problems. For example, learning from geological rock folding patterns analysis methods (Lisle, 1997; Li et al., 2010) defined the conjunction region of three gyral crests as a gyral hinge (denoted as 3-Hinges). Troubled by the formation mechanisms of 3-Hinges, Razavi et al. (2021) constructed a computational model of a growing brain and speculated that axonal wiring may be one of the most important contributors to 3-Hinges formation. The number, location, and shape of gyral hinges were used to quantitatively analyze the folding patterns of cerebral cortex (Nie et al., 2012; Ge et al., 2019; Huang et al., 2019). Gyral hinges receive an increasing attention not only because of their morphology, but also due to their importance in anatomy, axonal wiring diagram and brain functions: (1) they have thicker cortices (Li et al., 2010) and stronger axonal fiber connections (Ge et al., 2018); (2) they serve as the hubs of the cortico-cortical axonal fiber connective network (Zhang et al., 2020); and (3) they are more involved in global functional networks than other gyri (Zhang et al., 2020). According to recent studies, gyral hinges were suggested to serve as the anatomical landmarks, since corresponding gyral hinges across subjects were demonstrated to have unique and consistent structural connection patterns and brain function patterns (Zhang et al., 2020, 2022). In addition, some studies found that cortical folding pattern has significant differences among gender (Awate et al., 2010; Li et al., 2014). And these differences from the morphological structure of the cerebral cortex, especially the gyrus, may lead males and females to respond differently to the same cognitive activity (Charest et al., 2013; Hirjak et al., 2017).

Given the importance of gyral hinges, a more precise identification method is needed. In previous research, Yu et al. (2013) identified the gyral hinges by manual label. Chen et al. (2014) proposed a method based on energy minimization to identify the centroids of the gyral hinges with diffusion tensor imaging (DTI) derived fiber connectivity. Li et al. (2017) proposed an effective method for predicting the centroids of 3-Hinges based on DTI data using structural connection patterns and spatial distribution patterns. These methods significantly advanced the identification of 3-Hinges. However, they could not be easily generalized to the identification of 3-Hinges on large-scale cortical folding data since intensive manual intervention was involved. Subsequently, Chen et al. (2017) proposed a new representation of the cortical gyri pattern, named Gyral-net, which was automatically constructed as a gyral network. On this network, the nodes were automatically identified as gyral hinges, which are connected by gyral crests as edges (Chen et al., 2017; Zhang et al., 2018b). Despite the success of this automatic method, it takes a long time to only process the left or right brain of a single target at a time as the watershed algorithm and the tree marching algorithm are used such that it is hard to complete the identification task of gyral hinges on the dataset with a large amount of data.

Inspired by deep learning methods in many applications, Ge et al. (2019) applied convolutional neural network (CNN) to the cortical folding pattern recognition from functional magnetic resonance images (fMRI) to distinguish gyral hinges from other folding patterns. Although deep learning technique is promising in gyral hinge identification task due to its strength in latent feature exploration and utilization, the method in Ge et al. (2019) needs a precise cross-modality mapping to transfer the volumetric space of the fMRI data to the vertices on the cortical surface in T1-weighted MRI space, so did the method reported in Liu et al. (2022). Benefiting from the rich information of fMRI data, their work was influential on recognition of cortical folding pattern. However, instead of using the entire cortical fMRI data, they manually removed some data, according to cortical structure features. Furthermore, due to the huge variability of fMRI signals between individuals, both carried out their work at the individual level. In other words, a single model was trained for each subject, which consumed a lot of computing resources.

Therefore, this paper aims at developing a framework based on deep network models to realize the fast identification of cortical 3-Hinges simply from anatomic T1-weighted MRI and exploring whether there are structural differences on 3-Hinges among gender. The framework includes three major steps: Firstly, the morphological and structural features of the cerebral cortex are extracted from the reconstructed surface of the cerebral cortex. These features are then clustered into one feature vector per vertex using the K nearest neighbor algorithm. Secondly, based on this feature vector, cortical 3-Hinges folding regions are identified using a U-shaped neural network. Thirdly, the mean shift clustering algorithm is used to find the centroids of identified cortical 3-Hinges folding regions. Then, structural gender differences on 3-Hinges are discussed. The experimental results show that the proposed method can precisely recognize the locations of 3-Hinges and reveal the most contributive features to 3-Hinges identification, and there are significant differences in 3-Hinges morphological structure among adolescent gender.

2. Materials and methods

2.1. Overview

We propose a 3-Hinges locations identification algorithm based on a deep network trained on the morphological and structural features of the cerebral cortex. As shown in Figure 1, the algorithm framework includes three main steps: data preprocessing, identification of 3-Hinges regions (n is the number of vertices, k is the result of the K nearest neighbor algorithm, and m is the number of fusion data) and identification of 3-Hinges centroids. These steps will be detailed in the following subsections.

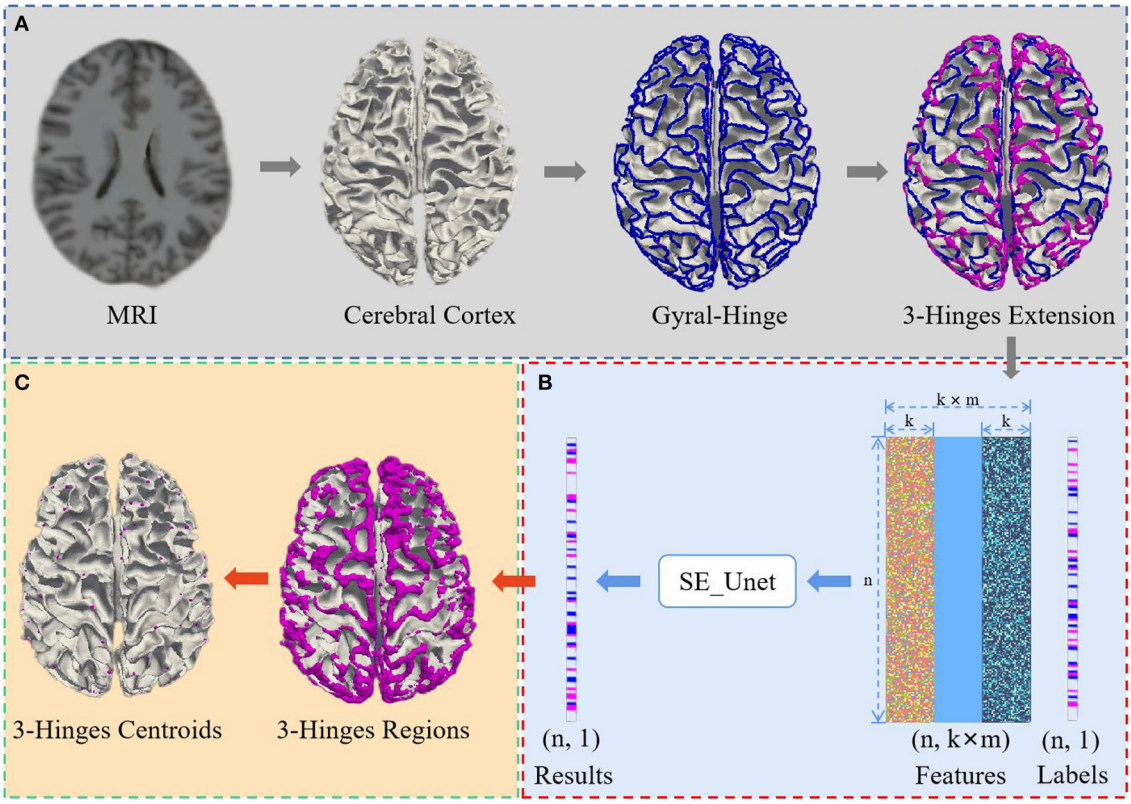


FIGURE 1
Overview of 3-Hinges locations identification framework. **(A)** Data preprocessing. **(B)** Identification of 3-Hinges regions. **(C)** Identification of 3-Hinges centroids.

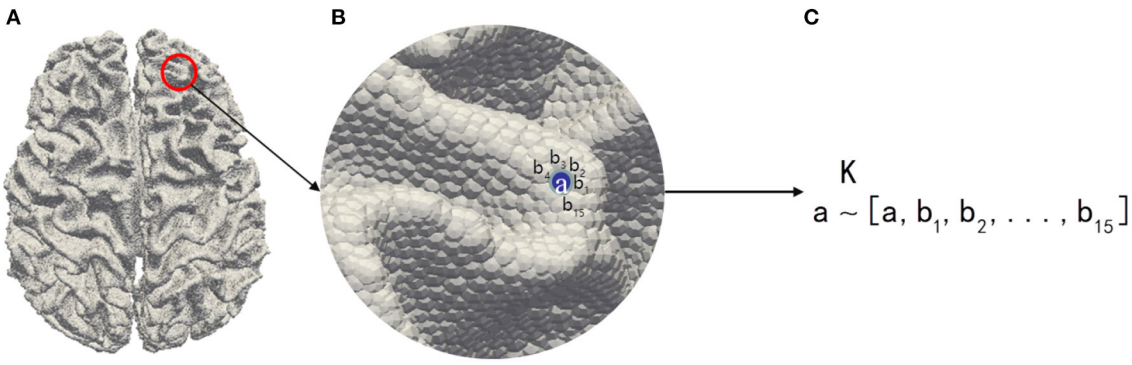


FIGURE 2
The illustration of feature preprocessing. **(A)** Cerebral cortex surface. **(B)** Zoom in view. **(C)** The feature vector [**a**,b_1,b_2,...,b_{15}] of the vertex **a** is aggregated by the K nearest neighbor algorithm.

2.2. Data preprocessing

2.2.1. Feature extraction and preprocessing

Considering the high ratio between the number of 3-Hinges centroids and the rest, we first use FreeSurfer (Fischl, 2012) for extracting features from the MRI reconstructed cortex to reduce the quantity ratio of non-3-Hinges to 3-Hinges. In this paper, we extract the morphological and structural features such as cortical thickness (thick), cortical surface area (area), cortical volume (vol), average curvature (curv) and sulcus (sulc) value to avoid using all the voxels in one brain as the input of the network model. In addition, because there are correlations among adjacent vertices on the cortex surface, we establish the spatial relationship between the scattered features with the K nearest neighbor algorithm (Cover and Hart, 1967; Pedregosa et al., 2011), and aggregate the morphological and structural features into a feature vector. For example, as shown in Figure 2, in our experiments, to each vertex a on the cortex surface, 15 vertices (b_1,b_2,...,b_{15}) in the

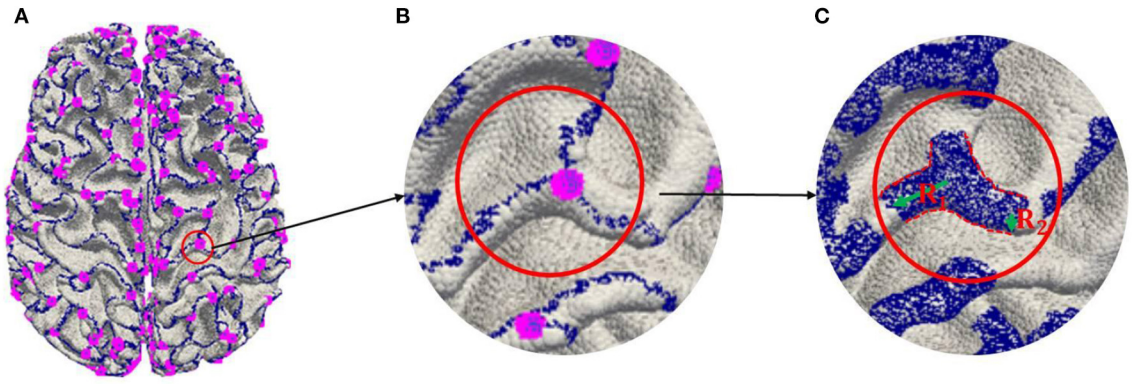


FIGURE 3
Labeling 3-Hinges regions. **(A)** Extracting 2-Hinges and 3-Hinges vertices by Gyral-net algorithm. **(B)** Zoom in 3-Hinges. **(C)** Expanding 2-Hinges and 3-Hinges regions.

neighborhood of the vertex a are selected as the single input feature experimentally.

2.2.2. 3-Hinges regions label

In order to further alleviate over-fitting in the training because of the imbalance between the number of 3-Hinges centroids and all the vertices of the cerebral cortex, three steps are involved in labeling 3-Hinges vertices.

(a) Extracting 2-Hinges and 3-Hinges vertices by the Gyral-net algorithm (the blue and the pink vertices are 2-Hinges and 3-Hinges vertices, respectively, as shown in Figure 3A. The readers can refer to Li et al. (2017) and Chen et al. (2017) about the detailed algorithm.

(b) Expanding 2-Hinges and 3-Hinges vertices into 3-Hinges region. As shown in Figure 3, we expand the vertices around the vertices generated by step (a). More specifically, two kinds of vertices are included in 3-Hinges region as shown in Figure 3C: i) cortex surface vertices in the spherical neighborhood within radius R_1 (empirically set to 6 mm) of 3-Hinges vertices; ii) the cortex surface vertices in the spherical neighborhood within radius R_2 (empirically set to 2 mm) of 2-Hinges vertices.

(c) Labeling 3-Hinges regions. We define the expanded region as 3-Hinges region shown as the blue region in Figure 3C. Each blue vertex is labeled as 1, and the rest is labeled as 0.

2.3. 3-Hinges regional identification

2.3.1. Single feature SE-Unet network framework

In this paper, we combine the U-shaped network structure (Ronneberger et al., 2015) and SE (Squeeze and Excitation) module (Hu et al., 2020) to design a SE-Unet network framework for the morphological and structural features, which are used to identify 3-Hinges regions automatically. As shown in Figure 4, the network framework is a symmetrical U-shaped network with two paths, encoding (left side) and decoding (right side), and a total of 5 layers. The encoding paths consists of the repeated application of two 3 × 3 convolutions (purple block), a SE module (yellow block, the architecture is shown in Figure 5), and a 2 × 2 max pooling operation with stride 2 for down sampling (green down-arrow). At each down sampling step, we double the number of feature channels. The decoding paths consists of an up sampling of the feature map followed by a 2 × 2 convolution that halves the number of feature channels (green up-arrow), skip connections (gray right-arrow) concatenation with the corresponding feature map from the encoding path, two 3 × 3 convolutions and a SE module. Specifically, each convolution is followed by a layer of batch normalization (BN) and a layer of ReLu activation function. Meanwhile, a dropout layer is put between the convolutional layers to alleviate over-fitting. The input data is converted to the range of [0, 1] by maximum and minimum normalization before fed into the first module composed of two layers of convolutional blocks and the SE module. In addition, the softmax function is applied before the output of the SE-Unet network. In order to facilitate network training, the dimension of the network input is designed to be 64 × 64 × 16 in our experiments. Besides, to reduce the number of learning-parameters and time consumption, the 2D convolution is utilized in the proposed network.

2.3.2. Multiple features fusion SE-Unet framework

For the extracted multiple feature vectors of the surface morphology and structure of the cerebral cortex, we design a multi-feature pre-fusion SE-Unet network framework to automatically extract 3-Hinges regions, as shown in Figure 6. The difference between this network structure and the single-feature SE-Unet network framework is that each feature in the input part of the network is first scaled by a convolutional block (including a 3 × 3 convolution layer, a layer of batch normalization (BN) and a layer of ReLu activation function), and then the scaled features are concatenated before being fed into the SE-Unet network.

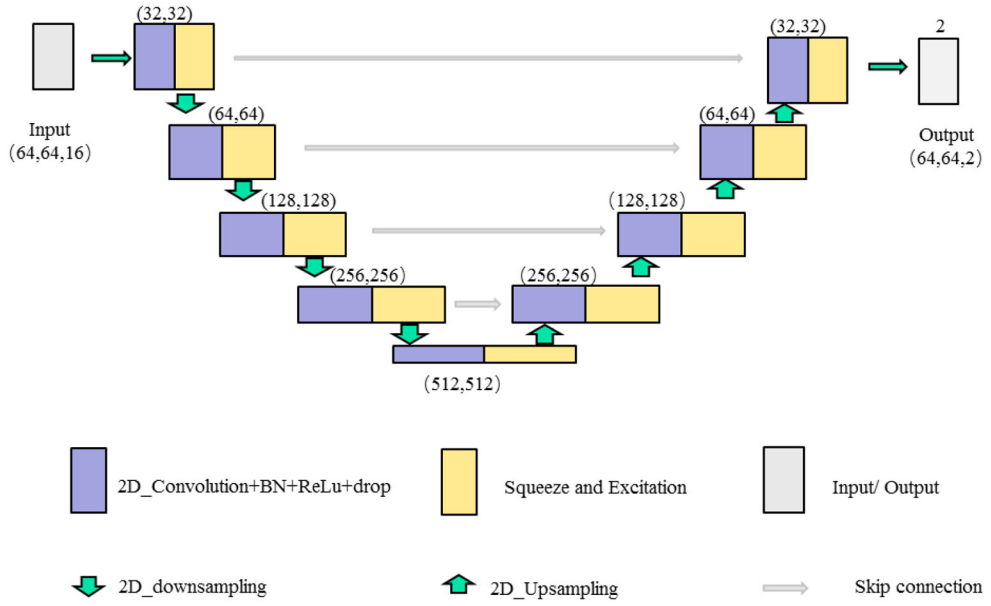


FIGURE 4
The illustration of single feature SE-Unet architecture.

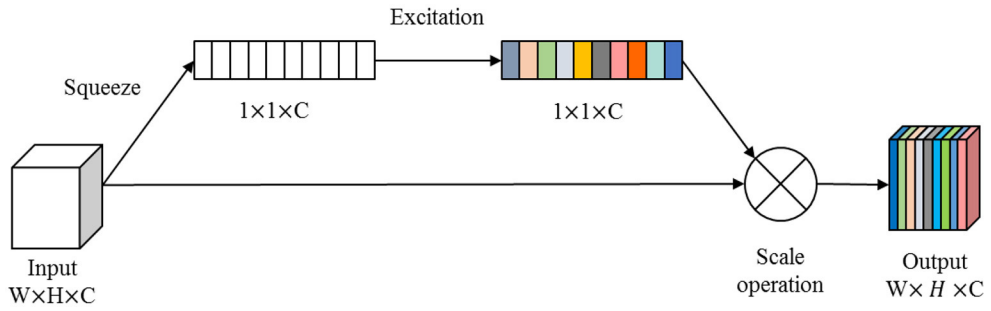


FIGURE 5
The architecture of the SE module.

2.4. 3-Hinges centroids identification

In order to identify the exact locations of 3-Hinges more precisely, we utilize the mean shift algorithm (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Collins, 2003) to cluster the centroids of 3-Hinges regions. Considering that the algorithm does not need to pre-define the number of cluster centers and that the number of 3-Hinges centroids is also unknown in advance, the algorithm can directly determine the cluster centroids based on the calculated offset mean vector.

Assuming that a certain 3-Hinges region X in the left/right brain hemisphere is composed of the 3-dimensional coordinate vector $X_i(i = 1, 2, ..., n)$, i.e., $X \in R^{n\times 3}$, the mean shift vector of $X_i(i \in \{1, 2, ..., n\})$ in the original mean shift vector can be calculated by the formula (1):

$$M_h(X_m) = \frac{1}{K}\sum_{X_i \in S_h}(X_i - X_m), \tag{1}$$

Where S_h is defined as the expression (Equation 2), h is the radius of 3-Hinges spherical region, and K is the number of coordinate vertices in 3-Hinges spherical region X.

$$S_h(X_m) = \{y : (y - X_m)^T(y - X_m) \le h\}. \tag{2}$$

However, the original mean shift algorithm assigns the same weight to each vertex in the region and regards them as the same importance. In fact, the closer the vertex is to the cluster center, the greater importance the vertex is to the cluster center. Therefore, the kernel function $G(\cdot)$ and weighted coefficients $w(\cdot)$ are introduced into the mean shift algorithm, and the formula (1) is modified as:

$$M_h(X_m) = \frac{\sum_{i=1}^{n} G_H(X_i - X_m)w(X_i)(X_i - X_m)}{\sum_{i=1}^{n} G_H(X_i - X_m)w(X_i)}, \tag{3}$$

Where $w(X_i) \ge 0$ is the weight corresponding to the coordinate vertex X_i according to the distance between X_i and X_m. $G_H(X_i - X_m)$

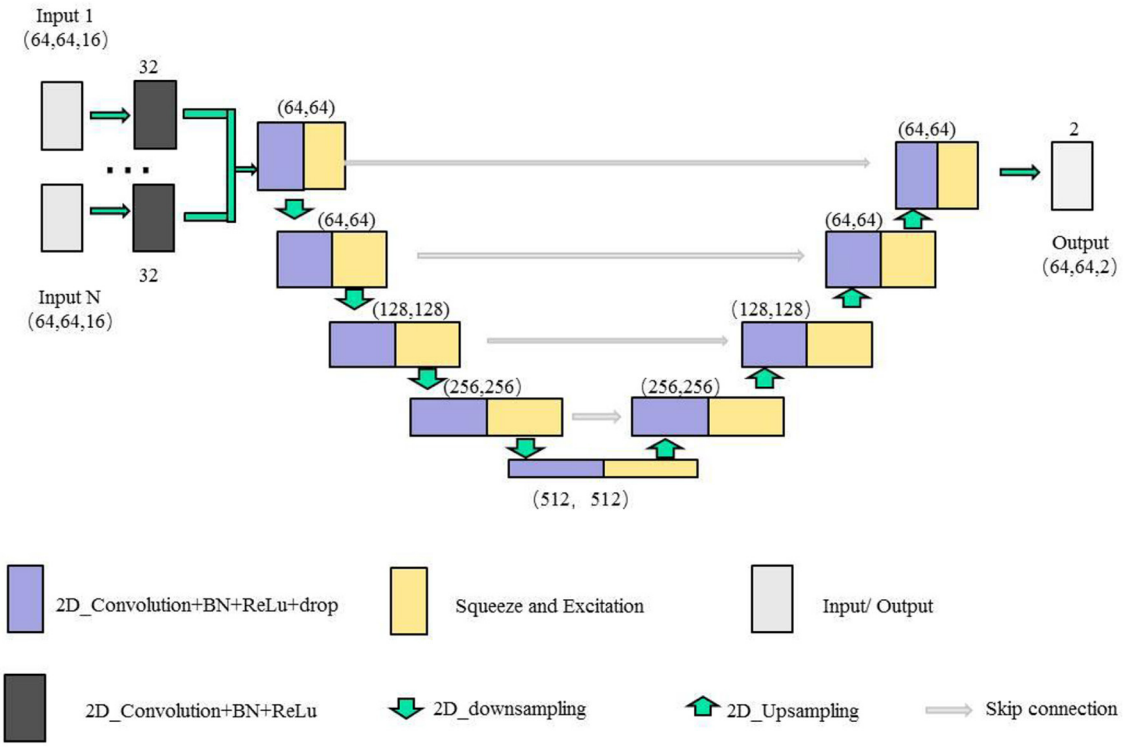


FIGURE 6
The illustration of the multiple features SE-Unet architecture.

is obtained by the expression:

$$\begin{aligned} & G_H(X_i - X_m) = |H|^{\frac{1}{2}} G_H(|H|^{\frac{1}{2}}(X_i - X_m)), \\ \textit{with} \quad & G_H(x) = -(\frac{1}{\sqrt{2\pi}s} e^{-\frac{x^2}{2s^2}})', s \in \textit{constant}, \end{aligned} \tag{4}$$

and H is a $d \times d$ bandwidth matrix, which can be the diagonal matrix $H = diag[h_1^2, ..., h_d^2]$ or the proportional unit matrix $H = h^2 I$. Considering that the later one only has one hyper-parameter h, we choose $H = h^2 I$ in the mean shift algorithm to facilitate the identification of 3-Hinges. After simplification, our final mean shift vector can be expressed as Equation (5):

$$M_h(X_m) = \frac{\sum_{i=1}^{n} G_H(\frac{X_i - X_m}{h}) w(X_i)(X_i - X_m)}{\sum_{i=1}^{n} G_H(\frac{X_i - X_m}{h}) w(X_i)}, \tag{5}$$

then, the 3-Hinges centroid is updated as $X_m = X_m + M_h(X_m)$.

3. Experimental results

In this section, we will introduce the data set, evaluation metrics, and network parameters. At the same time, we analyze the single feature and multiple combined features that are most relative to 3-Hinges. We also verify the generalization of the method on the adult data set. The code is available at https://github.com/GuardianTree/code.

3.1. Training

We evaluated our method on T1-weighted MR images from adolescent and adult data sets.

3.1.1. Data sets

The Adolescent MRI Data: In this study, the MRI from the Adolescent Brain Cognitive Development (ABCD) NIMH Data Archive (NDA) Study is used where all the subjects are between 9 and 10. Compared with infant brains, the brain at this age is considered to be relatively, with discriminative cortical folding patterns. The ABCD data set has been processed in accordance with the MRI preprocessing procedure mentioned by Jenkinson et al. (2002), Pfefferbaum et al. (2018), and Hagler et al. (2019). Limited by computational resources, we randomly select 1,000 brain MRI data from ABCD NDA Release 1.1. It is noted that the proposed method can be applied to many datasets including the above-mentioned datasets.

For the ABCD data set, there are approximately 330,000 vertices on the surface of the cerebral cortex of each sample. In order to facilitate the use of deep learning method, we will unify the features extracted from each sample to 331,776 (=64*64*81), that is, we add the morphological and structural features of the vertices that do not meet the requirements to 331,776 with a value of 0. After sampling and shape transformation, the features of each subject are divided into 81 blocks of size (64, 64, 16). Therefore,

there are 72,900 blocks in the training set and 8,100 blocks in the test set.

The Adult MRI data: In this experiment, the adult data set is the 1,200 data set released by the Human Connectome Project (HCP). The HCP data set contains images of a total of 1,200 normal young people aged 22–35. The detailed process of HCP data set parameters can be found in the processing of Van Essen et al. (2013). In order to verify the generalization of the adult data, 110 adults were selected from the HCP data set (http://www.humanconnectomeproject.org/data/).

For the HCP data set, there are approximately 360,000 vertices on the surface of the cerebral cortex of each subject. After the same processing as the ABCD data set, the extracted features are divided into 90 blocks, and the final HCP data set has 9,900 blocks with the size of (64, 64, 16).

3.2. Evaluation metrics

In this paper, three metrics are used to evaluate 3-Hinges regions identification performance in the experiment, i.e., Precision, Recall, and F1. In addition, in the process of identifying the locations of 3-Hinges centroids, the prediction error (PreE) calculated by the Euclidean distance between the predicted centroids and the labels is used as the evaluation metric. Lh-PreE, rh-PreE and mean-PreE represent the average values of the prediction error of 3-Hinges centroids on the left, right, and whole brain, respectively. The smaller the value of the PreE is, the closer the predicted 3-Hinges centroids locations are to the true 3-Hinges centroids locations.

3.3. Network parameters

In this experiment, we implement the SE Unet Network with the Keras framework, where the RMSprop optimizer Wilson et al. (2017) and Hinton et al. (2012) is used for optimization training. The initial learning rate is set as 0.05, which is decayed exponentially after each epoch. The batch size is set as 40, the epoch is set as 150, the convolution kernel size is set as 3×3, and the momentum parameter in the batch normalization layer is set as 0.6. The activation function layer is the ReLu function, the drop layer parameter is set as 0.2, the parameters in down-sampling and up-sampling are both set as 2 × 2. In order to obtain the true objective maximization of 3-Hinges regions, the Dice loss is selected as the training loss function.

3.4. 3-Hinges identification

3.4.1. Single feature result analysis

In the experiment, we first give the results of identifying 3-Hinges regions using the baseline U-net, and list the results using the proposed SE-Unet under different dimensionality reduction coefficients (r), which is a hyper-parameter in the SE module.

Then, based on the recognition of 3-Hinges regions, the mean shift clustering algorithm is used to identify the centroids of 3-Hinges regions. As shown in Table 1, when the hyper-parameter r is set to 24, the F1 score reaches 60.78, and the mean-pre of the predicted 3-Hinges centroids on the entire brain of all test set individuals is 5.56. Meanwhile, in the same experimental environment, the time consumption of our algorithm is about 4 min, which is far less than the Gyral-net method, indicating that our algorithm can identify the locations of 3-Hinges centroids more quickly.

Besides, we report the precision, recall and F1 under the other morphological and structural features of the cerebral cortex, such as cortical thickness, surface area, volume, average curvature and sulcus value, as shown in Table 2. We can see that under the same conditions, the sulc recognition results outperform those of other features. In addition, in the same experimental environment, compared with the Gyral-net method, the time required for our method is about 4 min, which are far less than Gyral-net method. As shown in Figure 7, we can observe that 3-Hinges regions identified by the sulcus value feature contains more 3-Hinges vertices which are close to the real 3-Hinges centroids. In some subjects, our predicted results are even more accurate than the labels annotated by Gyral-net such as those in (d-1) and (d-4) of Figure 7B.

3.4.2. Multiple features result analysis

Based on the experiment results of the single feature above, we try to improve 3-Hinges locations identification by fusing different features. In this section, we choose to use feature fusion in the early stage to explore the impact of fusion features on 3-Hinges locations identification, as shown in Table 3.

The result under the fusion of sulc+thick in 3-Hinges regions reaches 62.54, and the mean-PreE is only 5.23 mm. With sulc+curv the results are worse than that of a single sulc feature, which shows that the curv feature inhibits the sulc feature from identifying 3-Hinges locations. Similar conclusions are obtained from other feature combinations. We also get the optimal results with 3–5 features where it can be seen that more features do not improve the recognition results significantly, although the combination of sulc+thick+vol+area achieves better results at the cost of more time consumption. Some visualized results predicted by fusion of sulc feature and other structural features are shown in Supplementary Figures S1–S4. In general, our proposed method can predict some 3-Hinges points that are not labeled, such as a larger version of the left brain of individual a and d, and the right brain of individual b. Moreover, there are less 3-Hinges points, which are more likely to be representative in the same 3-Hinges region by using mean shift.

3.4.3. Correlation analysis with gender

We performed a correlation analysis between 3-Hinges cortical structural features classification accuracy and the subjects' gender, as shown in Table 4. In 100 test subjects, there are 51 females and 49 males. In single cortical structural feature tasks, there is not a significant correlation between 3-Hinges classification accuracy of one cortical structural feature and gender. But compared with the others, the result of cortical structural feature of the sulc have a

TABLE 1 The identification results of different methods.

Data	Methods		3-Hinges regions (%)			3-Hinges centroids (mm)			Time (min)
			Precision	Recall	F1	lh-PreE	rh-PreE	mean-PreE	
MRI	Gyral-net		-	-	-	-	-	-	82.30
sulc	Unet+ mean shift		56.23	65.71	60.56	5.58	5.63	5.60	4.06
sulc	SE_Unet+ mean shift	$r = 8$	55.32	67.28	60.67	5.55	5.63	5.59	4.05
		$r = 16$	55.47	67.12	60.70	5.56	5.61	5.58	4.07
		$r = 24$	55.74	66.93	60.78	5.52	5.60	5.56	4.06
		$r = 32$	56.49	64.35	60.12	5.54	5.59	5.56	4.05

TABLE 2 The identification results on different single features.

Data	Methods	3-Hinges regions (%)			3-Hinges centroids (mm)			Time (min)
		Precision	Recall	F1	lh-PreE	rh-PreE	mean-PreE	
MRI	Gyral-net	-	-	-	-	-	-	82.30
sulc	SE-Unet+ mean shift	55.74	66.93	60.78	5.52	5.60	5.56	4.06
curv		46.36	55.38	50.42	7.68	7.65	7.66	4.06
vol		47.84	55.21	51.23	6.92	6.86	6.89	4.09
area		44.60	43.49	44.01	8.46	8.46	8.45	4.07
thick		46.72	54.98	50.48	7.40	7.48	7.44	4.01

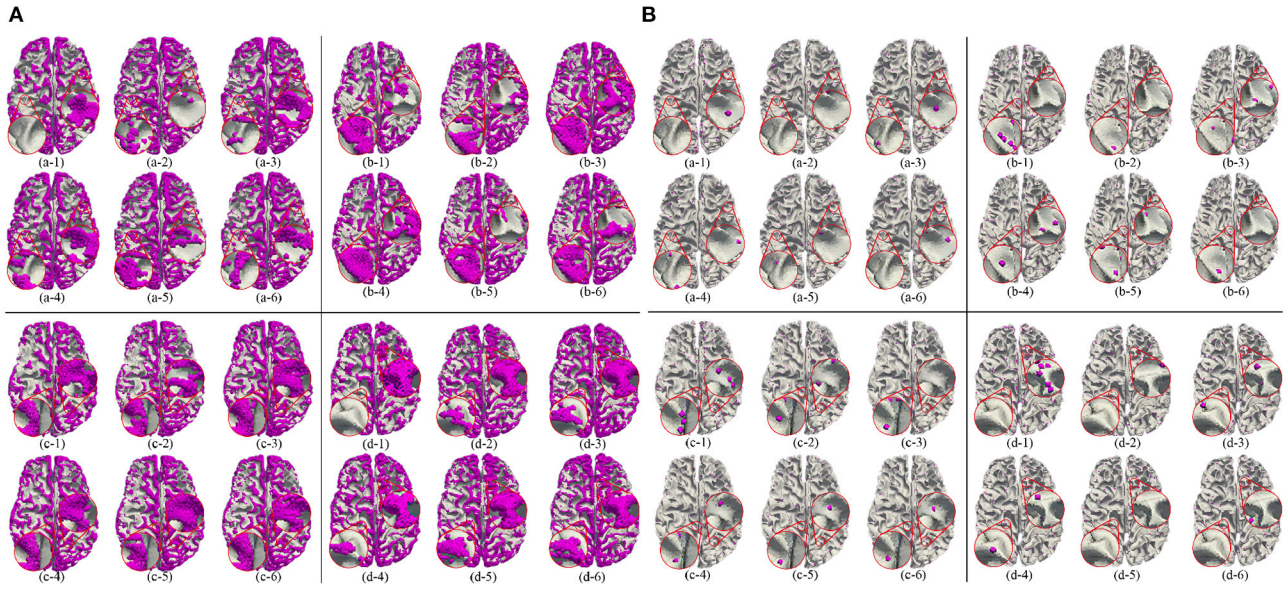


FIGURE 7
3-Hinges regions **(A)**/centroids **(B)** visualized results using different features. Letters (a–d) represent different individuals. Numbers (1–6) indicate label, area, curv, sulc, thickness, and volume, respectively.

TABLE 3 The identification results of multi-features fusion.

Data	Methods	3-Hinges regions (%)			3-Hinges centroids (mm)			Time (min)
		Precision	Recall	F1	lh-PreE	rh-PreE	mean-PreE	
MRI	Gyral-net	-	-	-	-	-	-	82.30
sulc+vol	SE-Unet +mean shift	55.91	70.33	62.21	5.26	5.25	5.25	4.08
sulc+thick		56.43	70.29	62.54	5.22	5.24	5.23	4.09
sulc+curv		50.25	74.87	59.84	5.94	5.93	5.94	4.07
sulc+area		56.13	69.30	61.91	5.52	5.51	5.51	4.09
vol+thick		49.21	56.80	52.68	6.82	6.84	6.83	4.11
vol+curv		43.97	42.70	41.80	7.37	7.33	7.35	4.15
vol+area		47.10	61.65	53.29	6.89	6.90	6.90	4.08
thick+curv		39.38	54.72	44.48	8.13	8.10	8.11	4.10
thick+area		47.71	55.78	51.30	7.49	7.47	7.48	4.13
curv+area		15.94	49.25	24.00	9.25	9.25	9.25	4.11
sulc+thick+vol		56.91	69.72	62.58	5.16	5.20	5.18	4.12
sulc+thick+area		56.59	69.23	62.15	5.21	5.22	5.21	4.11
sulc+vol+area		56.45	69.72	62.29	5.16	5.21	5.18	4.13
sulc+thick+vol+area		57.02	69.53	62.54	5.15	5.16	5.15	4.17
sulc+thick+vol+area+curv		51.70	74.02	60.35	5.55	5.65	5.60	4.32

TABLE 4 The correlation analysis between 3-Hinges regions identification accuracy and the gender in adolescents.

Data	*r*	*p*-value	Data	*r*	*p*-value
sulc	0.18	0.07	sulc+curv+vol	0.20	0.05*
curv	0.10	0.33	sulc+vol+area	0.24	0.02*
vol	0.11	0.27	sulc+thick+area	0.21	0.03*
area	0.16	0.10	sulc+thick+curv	0.21	0.04*
thick	0.15	0.14	sulc+thick+vol	0.19	0.06
sulc+area	0.19	0.06	sulc+vol+area+curv	0.21	0.04*
sulc+curv	0.21	0.03*	sulc+thick+area+curv	0.22	0.03*
sulc+vol	0.22	0.03*	sulc+thick+vol+area	0.18	0.08
sulc+thick	0.22	0.03*	sulc+thick+vol+curv	0.21	0.03*
sulc+area+curv	0.22	0.03*	sulc+thick+vol+area+curv	0.25	0.01**

The females and the males are labeled as 0 and 1, respectively. *represents p-value < 0.05, which means general significant correlation; **represents p-value < 0.01, which means extremely significant correlation.

closer association with gender ($r = 0.18$, $p = 0.07$). In two cortical structural features tasks, there is a significant correlation between 3-Hinges classification accuracy and gender ($r = 0.21$, $p = 0.03$ and $r = 0.22$, $p = 0.03$ for curv_sulc and sulc_thickness, respectively). In both three and four cortical structural features tasks, there are also significant correlations between 3-Hinges classification accuracies and gender ($r = 0.24$, $p = 0.02$ and $r = 0.22$, $p = 0.03$ for area_sulc_volume and area_curv_sulc_thickness, respectively). It is worth noting that in five cortical structural features tasks, there is the most significant correlation between 3-Hinges classification accuracy and gender ($r = 0.25$, $p = 0.01$). With the increasement of multiple cortical structural features, the correlation between 3-Hinges classification accuracy and gender becomes more and more significant, either the Pearson correlation coefficient or the p-value. Furthermore, all of the correlations are positive. It indicates that 3-Hinges structure of adolescent males is significantly different to that of females. Compared with other cortical folding regions, 3-Hinges regions are more prominent in males. This reslut is consistent with the previous study on gender differences in cerebral cortical folding patterns, in which the fraction of the cortical surface that was

TABLE 5 The identification results of HCP data set.

Data	Methods	3-Hinges regions (%)			3-Hinges centroids (mm)			Time (min)
		Precision	Recall	F1	lh-PreE	rh-PreE	mean-PreE	
MRI	Gyral-net	-	-	-	-	-	-	82.30
sulc	SE-Unet +mean shift	50.98	56.94	53.67	6.54	6.50	6.52	4.14
curv		46.16	59.01	51.65	8.13	8.25	8.19	4.18
vol		41.19	54.63	46.91	7.86	7.77	7.81	4.17
area		42.02	45.82	43.79	9.17	9.19	9.18	4.16
thick		42.16	49.50	45.48	8.09	8.22	8.16	4.13
sulc+thick		50.23	46.90	48.26	6.19	6.18	6.18	4.18
sulc+thick+vol		51.19	45.02	47.55	6.09	6.11	6.10	4.24
sulc+thick+vol+area		52.49	40.26	45.14	6.05	6.06	6.05	4.27
sulc+thick+vol+area+curv		47.76	56.18	50.73	6.52	6.47	6.49	4.32

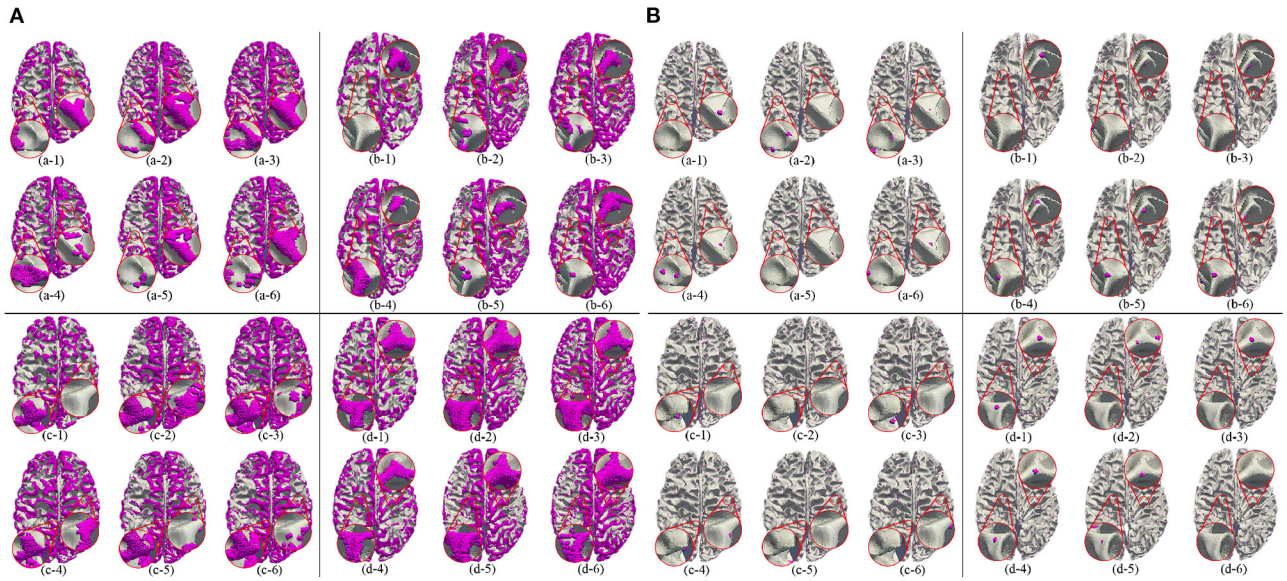


FIGURE 8
3-Hinges regions **(A)**/centroids **(B)** visualized results using different features on HCP. Letters (a–d) represent different individuals. Numbers (1–6) indicate label, area, curv, sulc, thickness, and volume, respectively.

convex (predominantly gyri including 3-Hinges) was significantly higher in males (Awate et al., 2009). In other words, structural roles that 3-Hinges within adolescent males and females plays do change remarkably.

3.5. Generalization

In this section, we test the adult data directly using the model trained on the ABCD data set. The results are shown in Table 5. It shows that we can get the consistent conclusions as the ABCD data set, although the accuracy is less than that of the ABCD data set. By analyzing and comparing the identified 3-Hinges regions and centroids, we find that on one hand, the adult brain is more mature than the adolescent brain, and its cerebral cortex folding is more complicated, which increases the difficulty of 3-Hinges' identification. On the other hand, the number of the vertices contained in the identified 3-Hinges regions is reduced, which results in less 3-Hinges centroids. However, as shown in Figure 8, the proposed method can still identify 3-Hinges points in some cases that are not correctly labeled by Gyral-net.

4. Discussion and conclusion

In this article, we propose a SE-Unet algorithm to identify 3-Hinges regions based on the extracted brain morphological

features. The algorithm first extracts the morphological and structural features of the brain, then utilizes the K nearest neighbor algorithm to establish the spatial index relationship between the scattered features and aggregates the extracted neighborhood features into a feature vector to improve the performance of the algorithm. At the same time, the deep U-shaped network structure and the squeeze excitation module are merged to learn the correlation of the channels in the feature vector, resulting in the automatic weight assignment of useful cortical structure feature channels. The cortical 3-Hinges regions can therefore be quickly identified. In addition, The mean shift algorithm is used to identify the centroids of the cortical 3-Hinges, considering that the cortical 3-Hinges is similar or identical in shape, which results in the inaccurate reflection of the cortical folding patterns. Through the comparative analysis of the experimental results of using a single feature and multiple features, we can conclude that the single sulc feature is sufficient to identify 3-Hinges. Meanwhile, the fusion of sulc, thickness, volume and area features can well identify 3-Hinges at the price of more time consumption. In consideration of the performance difference of identifying 3-Hinges between adolescent males and females, it is obvious that there are significant structural differences between males and females. In addition, we also carried out generalization verification on the adult dataset. Although our method improves the current Gyral-net to some extent, there are still room for improvement. We will aim for high accuracy prediction of the cortical 3-Hinges from both structural MRI and functional MRI.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The HCP study was ethically approved by the Washington University Institutional Review Board (IRB). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s), nor the minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

CC: supervision, writing–review, editing, validation, and project administration. LZ: methodology, writing–original draft, and coding. YL: writing–original draft, formal analysis, visualization, and coding. FH: supervision and project administration. XG: supervision and writing–review. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the Research Foundation of Education Department of Hunan Province of China (19A496, 21A0109, and 21B0172), the Natural Science Foundation of Hunan Province of China (2022JJ30571 and 2022JJ30552), Open Project of Key Laboratory of Medical Imaging and Artificial Intelligence of Hunan Province, Xiangnan University (YXZN2022003), and the National Natural Science Foundation of China (CN) (62272404 and 61972333).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1125666/full#supplementary-material

References

Awate, S. P., Yushkevich, P., Licht, D., and Gee, J. C. (2009). "Gender differences in cerebral cortical folding: multivariate complexity-shape analysis with insights into handling brain-volume differences," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2009, Vol. 5762*, eds D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor (Berlin; Heidelberg: Springer Berlin Heidelberg), 200–207.

Awate, S. P., Yushkevich, P. A., Song, Z., Licht, D. J., and Gee, J. C. (2010). Cerebral cortical folding analysis with multivariate modeling and testing: studies on gender differences and neonatal development. *Neuroimage* 53, 450–459. doi: 10.1016/j.neuroimage.2010.06.072

Charest, I., Pernet, C., Latinus, M., Crabbe, F., and Belin, P. (2013). Cerebral processing of voice gender studied using a continuous carryover fMRI design. *Cereb. Cortex* 23, 958–966. doi: 10.1093/cercor/bhs090

Chen, H., Li, Y., Ge, F., Li, G., Shen, D., and Liu, T. (2017). Gyral net: a new representation of cortical folding organization. *Med. Image Anal.* 42, 14–25. doi: 10.1016/j.media.2017.07.001

Chen, H., Yu, X., Jiang, X., Li, K., Li, L., Hu, X., et al. (2014). "Evolutionarily-preserved consistent gyral folding patterns across primate brains," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (Beijing: IEEE), 1218–1221.

Collins, R. T. (2003). "Mean-shift blob tracking through scale space," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, Vol. 2* (Madison, WI: IEEE), II-234.

Comaniciu, D., and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 603–619. doi: 10.1109/34.1000236

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Dubois, J., Benders, M., Borradori-Tolsa, C., Cachia, A., Lazeyras, F., Ha-Vinh Leuchter, R., et al. (2008). Primary cortical folding in the human newborn: an early marker of later functional development. *Brain* 131, 2028–2041. doi: 10.1093/brain/awn137

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

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T., et al. (2008). Cortical folding patterns and predicting cytoarchitecture. *Cereb. Cortex* 18, 1973–1980. doi: 10.1093/cercor/bhm225

Fukunaga, K., and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inf. Theory* 21, 32–40. doi: 10.1109/TIT.1975.1055330

Ge, F., Li, X., Razavi, M. J., Chen, H., Zhang, T., Zhang, S., et al. (2018). Denser growing fiber connections induce 3-hinge gyral folding. *Cereb. Cortex* 28, 1064–1075. doi: 10.1093/cercor/bhx227

Ge, F., Zhang, S., Huang, H., Jiang, X., Dong, Q., Guo, L., et al. (2019). "Exploring intrinsic functional differences of gyri, sulci and 2-hinge, 3-hinge joints on cerebral cortex," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Venice: IEEE), 1585–1589.

Hagler Jr, D. J., Hatton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., et al. (2019). Image processing and analysis methods for the adolescent brain cognitive development study. *Neuroimage* 202, 116091. doi: 10.1016/j.neuroimage.2019.116091

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Hirjak, D., Thomann, A. K., Kubera, K. M., Wolf, R. C., Jeung, H., Maier-Hein, K. H., et al. (2017). Cortical folding patterns are associated with impulsivity in healthy young adults. *Brain Imaging Behav.* 11, 1592–1603. doi: 10.1007/s11682-016-9618-2

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2011–2023. doi: 10.1109/TPAMI.2019.2913372

Huang, Y., He, Z., Liu, T., Guo, L., and Zhang, T. (2019). "Identification of abnormal cortical 3-hinge folding patterns on autism spectral brains," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy* (Springer International Publishing), 57–65. doi: 10.1007/978-3-030-33226-6_7

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

Jiang, X., Zhang, T., Zhang, S., Kendrick, K. M., and Liu, T. (2021). Fundamental functional differences between gyri and sulci: implications for brain function, cognition, and behavior. *Psychoradiology* 1, 23–41. doi: 10.1093/psyrad/kkab002

Jiang, X., Zhao, L., Liu, H., Guo, L., Kendrick, K. M., and Liu, T. (2018). A cortical folding pattern-guided model of intrinsic functional brain networks in emotion processing. *Front. Neurosci.* 12, 575. doi: 10.3389/fnins.2018.00575

Li, G., Wang, L., Shi, F., Lyall, A. E., Lin, W., Gilmore, J. H., et al. (2014). Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age. *J. Neurosci.* 34, 4228–4238. doi: 10.1523/JNEUROSCI.3976-13.2014

Li, K., Guo, L., Li, G., Nie, J., Faraco, C., Cui, G., et al. (2010). Gyral folding pattern analysis via surface profiling. *Neuroimage* 52, 1202–1214. doi: 10.1016/j.neuroimage.2010.04.263

Li, X., Zhang, T., Dong, Q., Zhang, S., Hu, X., Du, L., et al. (2017). "Predicting cortical 3-hinge locations via structural connective features," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC: IEEE), 533–537.

Lisle, R. (1997). Structural Geology of Rocks and Regions: Davis, G. H. and Reynolds, S. J. 1996. John Wiley and Sons, New York 2nd edition. *J. Struct. Geol.* 19, 752–753. doi: 10.1016/S0191-8141(97)85684-2

Liu, S., Ge, F., Zhao, L., Wang, T., Ni, D., and Liu, T. (2022). Nas-optimized topology-preserving transfer learning for differentiating cortical folding patterns. *Med. Image Anal.* 77, 102316. doi: 10.1016/j.media.2021.102316

Nie, J., Guo, L., Li, K., Wang, Y., Chen, G., Li, L., et al. (2012). Axonal fiber terminations concentrate on gyri. *Cereb. Cortex* 22, 2831–2839. doi: 10.1093/cercor/bhr361

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python, Vol. 12. *JMLR.org.* p. 2825–30. doi: 10.5555/1953048.2078195

Pfefferbaum, A., Kwon, D., Brumback, T., Thompson, W. K., Cummins, K., Tapert, S. F., et al. (2018). Altered brain developmental trajectories in adolescents after initiating drinking. *Am. J. Psychiatry* 175, 370–380. doi: 10.1176/appi.ajp.2017.17040469

Razavi, M. J., Liu, T., and Wang, X. (2021). Mechanism exploration of 3-Hinge gyral formation and pattern recognition. *Cereb. Cortex Commun.* 2, tgab044. doi: 10.1093/texcom/tgab044

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* (Springer), 234–241.

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

Van Essen, D. C. (1997). A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature* 385, 313–318. doi: 10.1038/385313a0

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. (2017). The marginal value of adaptive gradient methods in machine learning. *arXiv preprint* arXiv:1705.08292. doi: 10.48550/arXiv.1705.08292

Yu, X., Chen, H., Zhang, T., Hu, X., Guo, L., and Liu, T. (2013). "Joint analysis of gyral folding and fiber shape patterns," in *2013 IEEE 10th International Symposium on Biomedical Imaging* (San Francisco, CA: IEEE), 85–88.

Zhang, S., Wang, R., Han, Z., Yu, S., Gao, H., Jiang, X., et al. (2022). A dicccol-based k-nearest landmark detection method for identifying common and consistent 3-hinge gyral folding landmarks. *Chaos Solitons Fractals* 158, 112018. doi: 10.1016/j.chaos.2022.112018

Zhang, S., Zhang, T., Li, X., Guo, L., and Liu, T. (2018a). "Joint representation of cortical folding, structural connectivity and functional networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (Washington, DC: IEEE), 1–5.

Zhang, T., Chen, H., Razavi, M. J., Li, Y., Ge, F., Guo, L., et al. (2018b). Exploring 3-hinge gyral folding patterns among hcp q3 868 human subjects. *Hum. Brain Mapp.* 39, 4134–4149. doi: 10.1002/hbm.24237

Zhang, T., Li, X., Jiang, X., Ge, F., Zhang, S., Zhao, L., et al. (2020). Cortical 3-hinges could serve as hubs in cortico-cortical connective network. *Brain Imaging Behav.* 14, 2512–2529. doi: 10.1007/s11682-019-00204-6

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 April 2023
DOI 10.3389/fnins.2023.1115433

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China,
China

REVIEWED BY
Howard Goldman,
Cleveland Clinic,
United States
Thaddeus Brink,
Medtronic (United States),
United States

*CORRESPONDENCE
Limin Liao
✉ lmliao@263.net
Xing Li
✉ lxcpums@126.com

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

RECEIVED 04 December 2022
ACCEPTED 17 March 2023
PUBLISHED 05 April 2023

CITATION
Li X, Fang R, Liao L and Li X (2023) Real-time changes in brain activity during tibial nerve stimulation for overactive bladder: Evidence from functional near-infrared spectroscopy hype scanning.
Front. Neurosci. 17:1115433.
doi: 10.3389/fnins.2023.1115433

Real-time changes in brain activity during tibial nerve stimulation for overactive bladder: Evidence from functional near-infrared spectroscopy hype scanning

Xunhua Li[1,2], Rui Fang[3], Limin Liao[1,2,4]* and Xing Li[1,2]*

[1]Department of Urology, China Rehabilitation Research Center, School of Rehabilitation, Capital Medical University, Beijing, China, [2]University of Health and Rehabilitation Sciences, Qingdao, China, [3]Department of Occupational Therapy, China Rehabilitation Research Center, Beijing, China, [4]China Rehabilitation Science Institute, Beijing, China

Purpose: To use functional near-infrared spectroscopy (fNIRS) to identify changes in brain activity during tibial nerve stimulation (TNS) in patients with overactive bladder (OAB) responsive to therapy.

Methods: Eighteen patients with refractory idiopathic OAB patients were recruited consecutively for this pilot study. At baseline, all patients completed 3 days voiding diary, Quality-of-Life score, Perception-of-Bladder-Condition, and Overactive-Bladder-Symptom score. Then 4 region-of-interest (ROI) fNIRS scans with 3 blocks were conducted for each patient. The block design was used: 60s each for the task and rest periods and 3 to 5 repetitions of each period. A total of 360s of data were collected. During the task period, patients used transcutaneous tibial nerve stimulation (TTNS) of 20-Hz frequency and a 0.2-millisecond pulse width and 30-milliamp stimulatory current to complete the experiment. The initial scan was obtained with a sham stimulation with an empty bladder, and a second was obtained with a verum stimulation with an empty bladder. Patients were given water till strong desire to void, and the third fNIRS scan with a verum stimulation was performed. The patients then needed to urinate since they could not tolerate the SDV condition for a long time. After a period of rest, the patients then were given water until they exhibited SDV state. The fourth scan with sham fNIRS scan in the SDV state was performed. NIRS_KIT software was used to analyze prefrontal activity, corrected by false discovery rate (FDR, $p<0.05$). Statistical analyses were performed using GraphPad Prism software; $p<0.05$ was considered significant.

Results: TTNS treatment was successful in 16 OAB patients and unsuccessful in 2. The 3 days voiding diary, Quality-of-Life score, Perception-of-Bladder-Condition, and Overactive-Bladder-Symptom score were significantly improved after TNS in the successfully treated group but not in the unsuccessfully treated group. The dorsolateral prefrontal cortex (DLPFC) (BA 9, Chapters 25 and 26) and the frontopolar area (FA) (BA 10, Chapters 35, 45, and 46) were significantly activated during TNS treatment with an empty bladder rather than with an SDV. Compared with the successfully treated group, the unsuccessfully treated group did not achieve statistical significance with an empty bladder and an SDV state.

Conclusion: fNIRS confirms that TNS influences brain activity in patients with OAB who respond to therapy. That may be the central mechanism of action of TNS.

KEYWORDS

overactive bladder, tibial nerve stimulation, fNIRS, brain activity, central mechanism

Introduction

Overactive bladder (OAB) is characterized by urinary urgency, frequency, nocturia, and urgent incontinence in the absence of an infection or other evident disease by the International Continence Society (ICS) (Haylen et al., 2010). It affects numerous people, causing significant economic and quality of life problems (Stewart et al., 2003; Coyne et al., 2011; Reynolds et al., 2016). Treatment of OAB can be challenging, as many patients have persistent symptoms in spite of behavioral and oral pharmacologic therapies (Chancellor et al., 2014). Tibial nerve stimulation (TNS) is an alternative for those with OAB, and it comes in three forms: percutaneous (PTNS), implanted (ITNS), and transcutaneous (TTNS) (Schneider et al., 2015; Te Dorsthorst et al., 2020). Nonetheless, the precise mechanism of action in OAB therapy has yet to be determined.

Functional neuroimaging is useful in studying the brain micturition pathway (Fowler and Griffiths, 2010). According to functional neuroimaging studies, females with OAB had elevated afferent signaling to the cingulate, insular, and frontal cortices (Griffiths et al., 2005; Komesu et al., 2011). Several regions of the brain are essential for regular urination, and bladder filling also activates different brain regions (de Groat, 1998; Nardos et al., 2014; Griffiths, 2015). Studies using functional magnetic resonance imaging (fMRI) revealed higher activity in areas related with urine symptoms and urgency (Griffiths et al., 2007; Tadic et al., 2012). Functional near-infrared spectroscopy (fNIRS) has the benefits of noninvasive, portable, optic-based, and places little physical mobility limits to investigate the central micturition circuit (Duan et al., 2012; Geng et al., 2017). Furthermore, fNIRS has greater temporal resolution, can generate stable signals quicker, and can directly identify changes in oxyhemoglobin (HbO) signals in addition to deoxyhemoglobin (HbR) signals, making it superior to fMRI (Geng et al., 2017). Numerous fNIRS and fMRI studies have shown the accuracy and reproducibility of fNIRS signals, offering an evidential support for their use (Cui et al., 2011; Duan et al., 2012; Geng et al., 2017). In this study, we used fNIRS to study real-time brain activity during TNS treatment among OAB patients and explain the central mechanism of TNS.

Materials and methods

Patients

With Institutional Review Board approval (IRB:2021 N012), we recruited 18 women (mean age, 42.39 ± 19.72 years) with refractory idiopathic OAB who chose TTNS. The inclusion criteria were as follows: age 18 to 75 years, 72 h of recording urination with at least 8 voids every day and 7 days of abstaining from anticholinergic and β3 adrenergic receptor agonist prior to TTNS. Drug usage was unchanged throughout therapy. Patients with untreated symptoms of urinary tract infection, bladder tumor, or urinary stones were ineligible, as were those who were pregnant, had a pacemaker or implanted defibrillator, had combined renal insufficiency, Parkinson's disease, complete spinal cord injury, mental illness that prevented them from cooperating with doctors, skin lesions at the treatment place, and had participated in other drug or device clinical trials within 1 month prior to enrollment.

Stimulation procedures

Evaluations were not carried out when the subjects were having their periods. At the beginning of the study, every patient recorded their voiding diary for 72 h, received a score on their Quality of Life (QoL), assessed their Perception of Bladder Condition (PPBC), and completed an Overactive Bladder Symptom score (OABSS). If patients met the inclusion exclusion criteria, we then conducted the fNIRS trial on them. Patients were instructed on how to use the stimulator after the experiment, and they then went home to stimulate themselves. The stimulate parameters was as follows: 20-Hz frequency and a 0.2-millisecond pulse width and 30-milliamp stimulatory current. Patients performed TNS 1 h per day for 30 days and then returned to our facility to follow up and complete a 72-h voiding diary prior to the follow-up day as well as a Qol score, PPBC score, and OABSS. Clinical treatment success was characterized as either a decrease of daily frequency voids of at least 30% or a reduction of urgency voids of at least 50% (Cava and Orlin, 2022). Region-of-interest (ROI) fNIRS scans were showed in Table 1.

The fNIRS experiment flow was as follows: upon accessing the research facility, subjects were informed a description of the experiment, given a permission form, and instructed to take a seat. In a line with the tibial nerve, the 2 mucilaginous electrodes of the stimulator were inserted roughly three fingers above the medial malleolus. Patients had fNIRS electrodes placed on their foreheads and then they closed their eyes in a darkened environment. A total of 4 fNIRS scans containing 3 blocks each were completed for each patient. The block design was used: 60 s each for the task and rest periods, and 3 to 5 repetitions of each period. Fifteen seconds of baseline resting data were added before the block to ensure the steady state of the fNIRS signal, and a total of 360 s of data were collected.

During the task period, patients used TTNS (General Stim, Inc., Hangzhou, Zhejiang, China) on the right lower limb with parameters of 20 Hz frequency, a 0.2-millsecond pulse width, and a 30-milliamp (mA) stimulatory current. The initial scan was obtained with sham stimulation (using the same TTNS device and parameters but the power of the device was off which inducing no stimulation effects) with an empty bladder and a second time with verum stimulation with an empty bladder. The third fNIRS scan was conducted on the subjects after they were given water until they exhibited a strong desire to void (SDV) without being concerned about leaking. Because OAB patients cannot maintain urine storage for a long period with SDV, patients needed to void after the third fNIRS scan. After a period of rest, the patients then were given water until they exhibited SDV state. The fourth scan with sham fNIRS scan in the SDV state was performed (Figure 1). The block design provides many advantages, including the reduction of the need for human involvement and the suppression of oscillations in data that are not relevant (Sato et al., 2007).

fNIRS equipment

To monitor the variations in HbO and HbR in the venous blood of the cortex cortical areas, a two-channel fNIRS topography apparatus (Shimadzu Co.) was utilized. Light-NIRS is capable of capturing hemodynamic responses by concurrently irradiating near-infrared

TABLE 1 Channel locations for the fNIRS cap.

Ch	MNI coordinates (*x y z*)			BA	Brain area	Probability
1	34.84	−9.26	69.78	6	Pre-motor and supplementary motor cortex	0.9
2	−42.88	−10.92	63.07	3	Somatosensory cortex	1
3	47.77	−13.74	61.09	3	Somatosensory cortex	1
4	38.29	15.42	58.52	6	Pre-motor and supplementary motor cortex	0.9
5	−43.63	13.35	53.61	6	Pre-motor and supplementary motor cortex	0.9
6	−53.42	−16.59	56.15	2	Somatosensory cortex	1
7	56.89	−23.36	52.98	40	Supramarginal gyrus	0.71
8	50.52	5.28	50.87	6	pre-motor and supplementary motor cortex	0.9
9	37.51	29.64	49.48	8	Includes frontal eye field	0.61
10	−42.72	27.26	44.72	8	Includes frontal eye field	0.61
11	−55.13	2.24	44.44	6	Pre-motor and supplementary motor cortex	0.9
12	−61.04	−26.06	46.01	40	Supramarginal gyrus	0.71
13	64.65	−25.85	41.39	40	Supramarginal gyrus	0.71
14	57.96	4.06	40.64	6	Pre-motor and supplementary motor cortex	0.9
15	46.1	30.82	39.25	8	Includes frontal eye field	0.61
16	28.13	50.72	36.07	8	Includes frontal eye field	0.61
17	−36.8	46.73	31.46	9	Dorsolateral prefrontal cortex	0.79
18	−51.78	24.48	33.06	9	Dorsolateral prefrontal cortex	0.79
19	−61.63	−2.81	34.68	6	Pre-motor and supplementary motor cortex	0.9
20	−65.04	−29.62	38.38	40	Supramarginal gyrus	0.71
21	63.37	−1.67	30.79	6	Pre-motor and supplementary motor cortex	0.9
22	53.06	28.62	29.94	9	Dorsolateral prefrontal cortex	0.79
23	40.07	49.89	25.65	9	Dorsolateral prefrontal cortex	0.79
24	20.2	63.73	23.77	9	Dorsolateral prefrontal cortex	0.79
25	−7.92	66.05	23.29	9	Dorsolateral prefrontal cortex	0.79
26	−26.4	60.8	21.4	9	Dorsolateral prefrontal cortex	0.79
27	−44.66	45.28	21.7	46	Dorsolateral prefrontal cortex	0.61
28	−55.98	21.56	24.68	9	Dorsolateral prefrontal cortex	0.79
29	−65.22	−7.35	27.28	4	Primary motor cortex	0.98
30	67.04	−9.3	17.27	3	somatosensory cortex	1
31	59.3	19.86	19.5	9	Dorsolateral prefrontal cortex	0.79
32	48.7	46.53	12.36	46	Dorsolateral prefrontal cortex	0.61
33	29.74	63.63	13.27	10	Frontopolar area	0.92
34	9.37	70.37	12.07	10	Frontopolar area	0.92
35	−16.19	68.4	13.62	10	Frontopolar area	0.92
36	−38.43	58.91	9.29	10	Frontopolar area	0.92
37	−52.28	39.05	9.49	46	Dorsolateral prefrontal cortex	0.61
38	−60.88	11.92	14.38	44	Pars opercularis Broca's area	0.73
39	−66.83	−14.27	15.45	43	Subcentral area	0.68
40	62.97	7.33	8.98	6	Pre-motor and supplementary motor cortex	0.9
41	53.89	39.99	4.04	46	Dorsolateral prefrontal cortex	0.61
42	40.63	60.05	0.58	10	Frontopolar area	0.92
43	19.42	70.09	2.84	10	Frontopolar area	0.92
44	−8.28	70.72	0.82	10	Frontopolar area	0.92

(Continued)

TABLE 1 (Continued)

Ch	MNI coordinates (*x y z*)			BA	Brain area	Probability
45	−28.68	64.9	0.19	10	Frontopolar area	0.92
46	−44.19	54.77	−2.08	10	Frontopolar area	0.92
47	−54.88	34.76	0.43	45	Pars triangularis	0.7
48	−63.53	−1.4	−2.08	22	Superior temporal gyrus	0.46

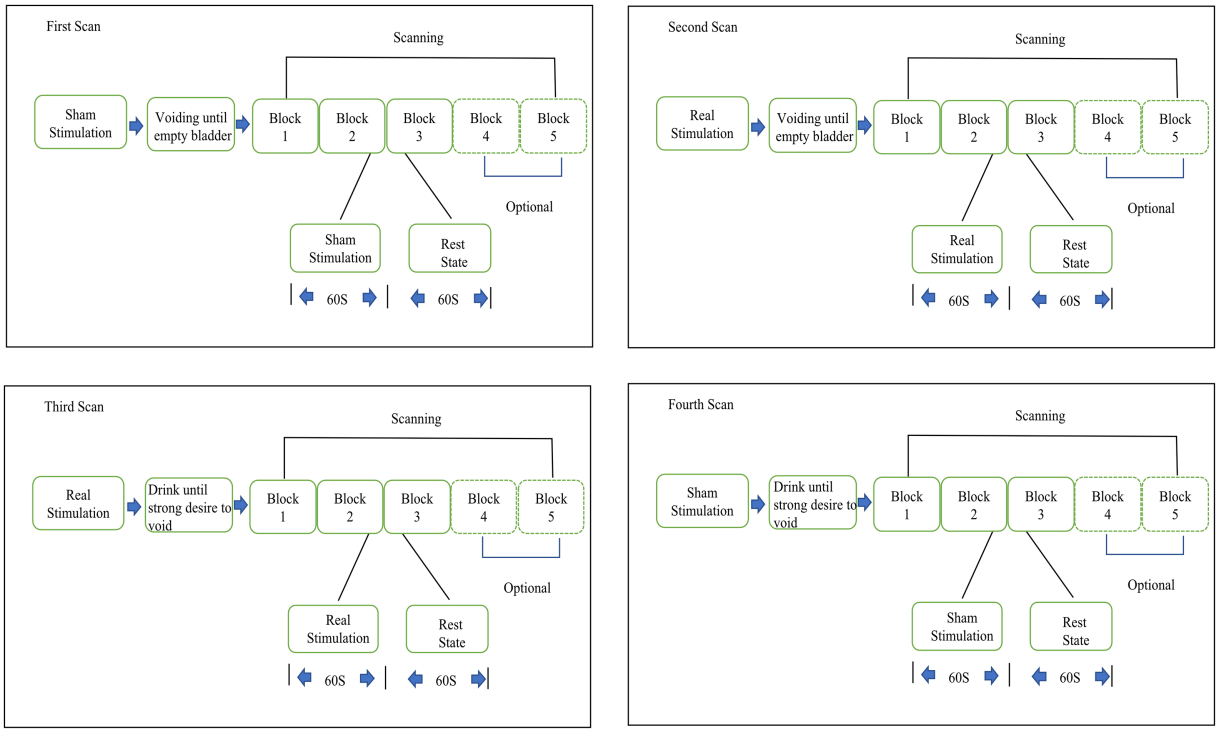


FIGURE 1
Block diagram of the fNIRS experimental design.

light in three wavelengths (780, 805, and 830 nanometers) using optical cables. The probe system, consisting of a skull cap with 16 near-infrared light emitters, 16 detectors, and 48 channels, was placed on the frontal lobe, with the lowest probes located along the Fp1-Fp2 line (Okada and Delpy, 2003)(Figure 2). A 3D digitizer (Patriot; Polhemus) was used to generate the position information of total circuits and evaluated utilizing NIRS_SPM to get the Montreal Neurological Institute (MNI) coordinates and the possibility of connected brain areas in the Brodmann area (BA) atlas (Jiang et al., 2020; Xu et al., 2020). Channel location details are shown in Table 1.

fNIRS data analysis

A MATLAB toolbox (Hou et al., 2021) was used to perform data preprocessing and visualize the results. To guarantee a steady signal, the fNIRS data were trimmed by the initial and final 15 s. We used a first-order detrend to get rid of the sluggish time-based fluctuations (Racz et al., 2018). The temporal derivative distribution repair method was used for motion correction (Fishburn et al., 2019). In addition, artifacts were removed by band pass filter limiting the data between 0.008 and 0.08 Hz (Bulgarelli et al., 2020). In this investigation, we focused only on variations in HbO since that signal has been shown to be more sensitive than HbR in detecting differences in regional cerebral blood circulation (Fu et al., 2014). After fNIRS data preprocessing, the individual-level analysis may be performed using the mass univariate statistical approach based on GLMs. For the statistical analysis, the steps listed below were used. To begin, creating a GLM that models the observed hemodynamic signal as a linear mixture of target regressors, unwanted variables, and an error term. Constructing the reference time series representation from task variables using the canonical hemodynamic response function defined in SPM is required for GLM definition. Then, the estimation of GLM parameters on a channel-by-channel basis, which fined the activation beta value for each experimental condition. In the end, utilizing contrast vectors from the pre- and post-stimulus as the input for subsequent group-level inference, the condition-wise effects were calculated. Paired t-test was used for the group-level analyses, corrected by false discovery rate (FDR, $p < 0.05$) (Hou et al., 2021).

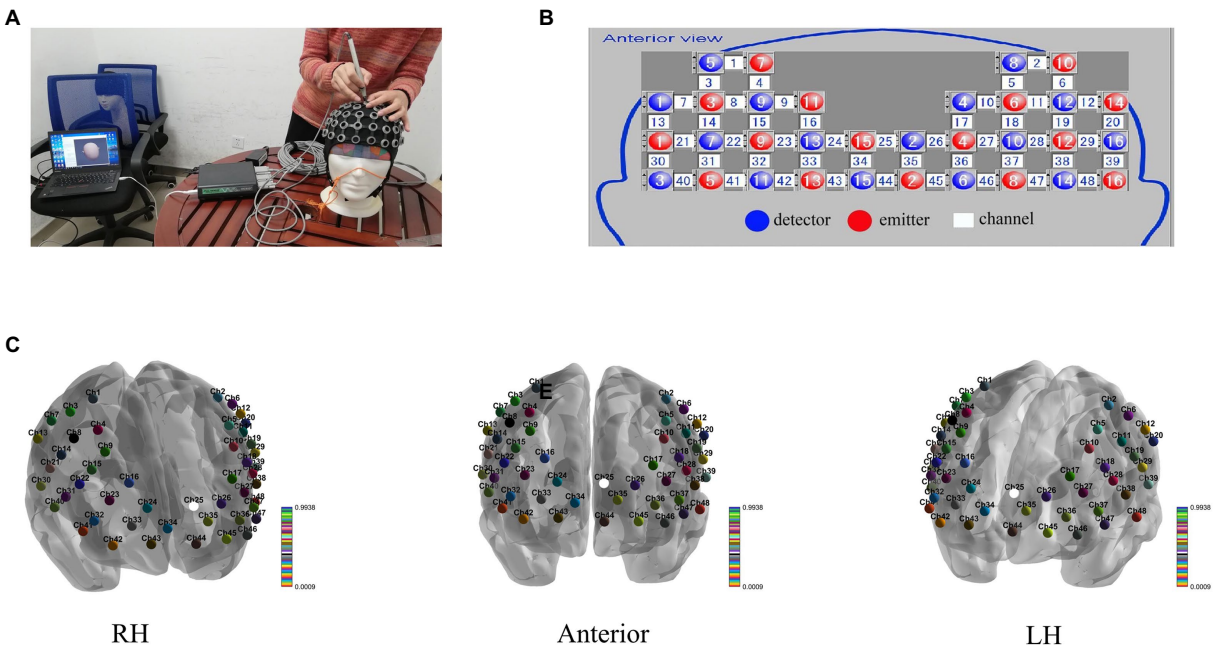


FIGURE 2
(A) Transcranial mapping navigation to locate areas of interest, **(B)** the sensor array, and **(C)** the 48 channels' 3D MNI coordinates.

Statistical analyses

We used GraphPad Prism software to conduct statistical analyses. Descriptive data were descripted as mean ± SD or median (25th to 75th percentile) in accordance with the assumption of the normality of the data. Student's t-test or Wilcoxon test was done on paired continuous variables according to the kind of distribution. $p < 0.05$ was regarded statistically significant.

Results

Eighteen right-handed women with OAB who elected TTNS treatment were included in our research. Prior to the experiment, none of the patients received TTNS. Sixteen patients were treated successfully, while two were unsuccessfully treated. Table 2 shows baseline statistics of successfully treated patients. Among the patients, 4 had OAB-wet and 11 had nocturia.

Comparison of voiding data before and after TTNS treatment

The clinical parameters showed varying degrees of substantial improvement relative to pretreatment levels (Table 3). The average daily number of micturition, incontinence episodes, and urgency score were decreased from 13.40 ± 2.23 to 7.79 ± 1.22, 6.50 (0.75 to 13.75) to 4.17 (0.00 to 8.59), and 3.62 (0.90 to 3.92) to 2.00 (0.00 to 2.70), respectively. The mean voiding volume was increased from 125.80 ± 33.42 mL to 149.00 ± 36.74 mL. The OABSS, QoL, and PPBC were reduced from 6.06 ± 2.52 to 3.94 ± 2.86, 4.63 ± 0.96 to 2.56 ± 1.79, and 4.50 ± 1.10 to 2.88 ± 1.54, respectively.

TABLE 2 Baseline demographic and clinical characteristics of successfully treated patients.

	OAB (n =16)
Age, years	36.25 ± 16.04
BMI, kg/m^2	22.34 ± 3.02
OAB Type	
OAB-Dry	12 (75%)
OAB-Wet	4 (25%)
Duration of OAB symptoms, years	3.97 ± 2.18
Handedness	Right-handed

Comparison of fNIRS data between sham stimulation and verum stimulation in empty bladder and SDV in successfully treated group

During the sham stimulation condition, patients with an empty bladder showed no significant changes in any brain regions between the stimulation and rest states. The T-values between the two states are shown Figure 3A. However, in the verum stimulation state, there was significant activation in some brain areas between the stimulation and rest states, such as dorsolateral prefrontal cortex (DLPFC) (BA 9, Chapters 25 and 26), and the frontopolar area (FA) (BA 10, Chapters 35, 45 and 46). The T-values between the two states are shown in Figure 3B. In the SDV state, there were no significant changes in any brain areas both in verum stimulation (Figure 3C) and sham stimulation conditions (Figure 3D).

TABLE 3 Clinical parameters before and at completion of TNS treatment.

Parameters	Pre-treatment	Post-treatment	*p*-value
Micturition frequency daily	13.40 ± 2.23	7.79 ± 1.22	<0.05
Mean voiding volume (mL)	125.80 ± 33.42	149.00 ± 36.74	<0.05
Number of incontinence episodes per day	6.50 (0.75–13.75)	4.17 (0.00–8.59)	<0.05
Number of Nocturia	1.85 ± 0.85	1.09 ± 0.54	<0.05
Urgency Score	3.62 (0.90–3.92)	2.00 (0.00–2.70)	<0.05
OABSS	6.06 ± 2.52	3.94 ± 2.86	<0.05
QoL	4.63 ± 0.96	2.56 ± 1.79	<0.05
PPBC	4.50 ± 1.10	2.88 ± 1.54	<0.05
PVR	<10 mL	<10 mL	

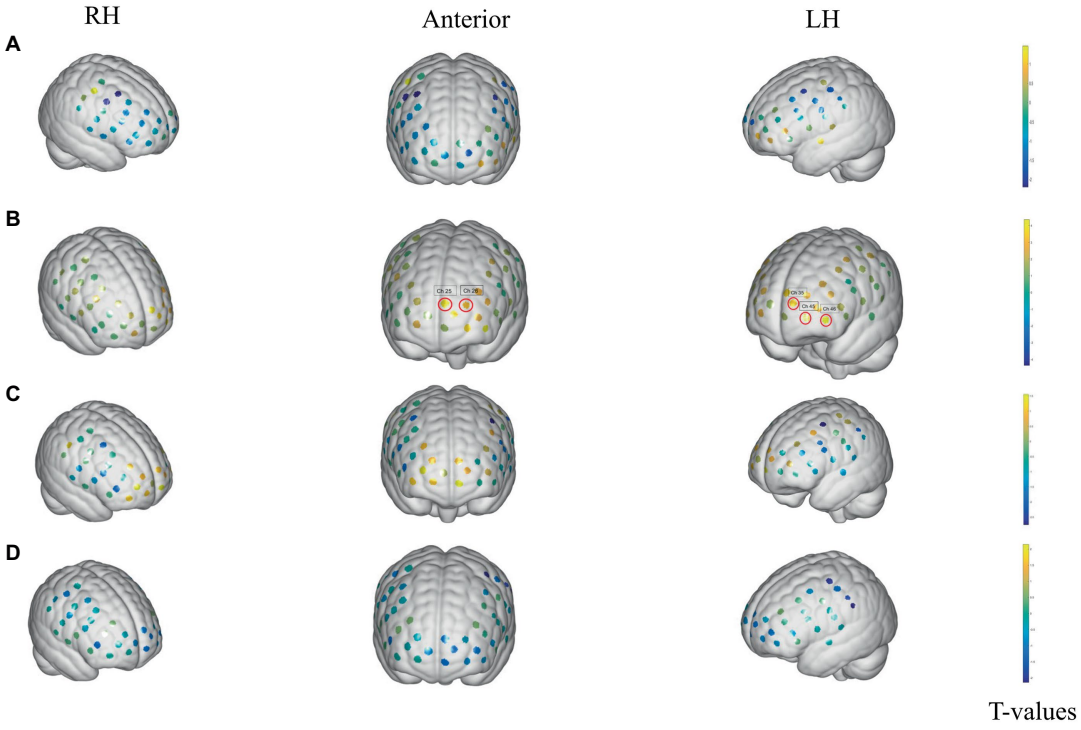


FIGURE 3
Activation changes of ROI fNIRS data in the empty-bladder and SDV states in successfully treated group. **(A)** Activation of sham stimulation state with empty bladder. **(B)** Activation of verum stimulation with empty bladder. **(C)** Activation of verum stimulation in SDV state. **(D)** Activation of sham stimulation in SDV state. The channels are denoted by the dots. Significance in activation variations is shown by red loops in front of channels ($p<0.05$, FDR corrected). The colored bar reflects group-level *T*-values. *T* The cool hue represents deactivation, whereas the bright color represents activation.

Comparison of fNIRS data between sham stimulation and verum stimulation in empty-bladder and SDV states in the unsuccessfully treated group

The unsuccessfully treated patients with both an empty bladder and SDV state achieve no significant changes in any ROIs both in the sham stimulation state and the verum stimulation state. The sample size of the unsuccessfully treated group was only two, and a larger sample is needed to verify the results.

Discussion

This is the very first prospective research to evaluate the central TNS mechanism in OAB patients utilizing fNIRS. We found regional brain activation with an empty bladder after successful TTNS in women with OAB. Areas activated included the DLPFC, and FA during TTNS. Furthermore, patterns of brain activity differed between women who responded to TTNS and those who were unsuccessfully treated. Different functional neuroimaging devices have been used to explore brain function during urination for some time. As early as

1996, a study based on CT and MRI found that subjects with frontal-lobe lesions showed detrusor hyperreflexia and unrestrained sphincter slackness, resulting in lower urinary tract symptoms (Sakakibara et al., 1996). SPECT and PET technologies have been steadily utilized to neuroimaging during the last several decades due to the fast growth of functional brain imaging technologies (Fukuyama et al., 1996; Blok et al., 1997, 1998; Nour et al., 2000). After that, fMRI and fNIRS were used to investigate the centralized bladder control mechanism that had been predicted. fMRI measures HbR paramagnetism and has exceptional temporal and spatial resolution (Kitta et al., 2015), whereas fNIRS is based on HbO and HbR absorption of near-infrared light and has the benefits of mobility, outstanding temporal resolution, and convenient for clinical use (Jobsis, 1977).

The mechanism of brain function in urination is not still completely understood. Previous studies suggested a functional paradigm for bladder control, including the brain areas such as thalamus, insula, prefrontal cortex (PFC), and periaqueductal gray (PAG) (Griffiths et al., 2005; de Groat et al., 2015). The DLPFC is primarily responsible for executive functions, including the consolidation of information from multiple senses, preservation of focus, and management of goal-directed activity. According to a fNIRS research, the bilateral DLPFC was highly active in the SDV condition, and the greater the urge to urinate, the greater the bilateral DLPFC activation (Matsumoto et al., 2011). Our earlier work demonstrated aberrant DLPFC deactivation in OAB patients, which may relieve DLPFC inhibition on the voiding reflex (Pang et al., 2022).

TNS is a crucial component in the treatment of OAB since it is both effective and less invasive. Previous investigations have offered clues on the potential mechanisms include inhibition of threshold afferent nerve activity (Choudhary et al., 2016), increasing endogenous opioid peptide levels in the central nervous system (Matsuta et al., 2013), and inducing bladder inhibition through cerebral cortex network reconstruction (Finazzi-Agro et al., 2009). During TTNS, brain areas such as the DLPFC (BA 9, Chapters 25 and 26) and the FA (BA 10, Chapters 35, 45 and 46) were activated in the current study. It seems that TTNS could help relieve OAB symptoms by activating brain areas crucial to the voiding reflex. Griffiths et al. (2005) found that OAB patients showed significantly weaker responses to infusion than healthy patients especially in the anterior insula. When the bladder was completely filled, the infusion elicited heightened reactions throughout most of the brain. Still, the reaction in the orbitofrontal cortex was much weaker than it was in individuals with strong control. In this study, OAB patients with an empty bladder achieved significant activation in the BA 9 to 10 areas, compared with the stimulation and rest states. However, they did not achieve activation when patients' bladders were full. This may be because these brain areas are more activated with a full than an empty bladder, and the difference between activation generated by stimulation and that produced by bladder filling is reduced. Our findings suggest that TNS's potential primary mechanism for OAB is the normalization of the voiding reflex and the restoration of DLPFC, and FA activation.

This study has limitations. Due to the insufficient size of the patient sample, the findings were not adjusted for the full complement of channels. Furthermore, fNIRS could not monitor the activation alteration of the whole brain cortex and deep brain structures due to the limitations of the detection range imposed by the number and penetration depth of probes. Improved fNIRS technology and analytical techniques may eventually solve this problem.

Conclusion

TNS has an effect on the brain function of OAB patients who show a clinical response to the treatment. To some extent, it may be how TNS works to alleviate OAB. In subsequent research, fMRI may be used to analyze changes in brain activity associated with clinical responses to medication.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the China Rehabilitation Research Center review board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LML designed the study. XHL, RF, and XL conducted the research. XHL wrote the manuscript. LML and XL contributed to significant modifications of vital knowledge content. The final version has been authorized by all writers, who accept responsibility for all parts of the work. The final text was reviewed and approved by all writers. All authors contributed to the article and approved the submitted version.

Funding

The Ministry of Science and Technology of the People's Republic of China supported this research (2018YFC2002203). The funders had no part in the original study concept, information collection and analysis, publication decision, or manuscript writing.

Conflict of interest

This study was funded by 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

References

Blok, B. F., Sturms, L. M., and Holstege, G. (1998). Brain activation during micturition in women. *Brain* 121, 2033–2042. doi: 10.1093/brain/121.11.2033

Blok, B. F., Willemsen, A. T., and Holstege, G. (1997). A PET study on brain control of micturition in humans. *Brain* 120, 111–121. doi: 10.1093/brain/120.1.111

Bulgarelli, C., Ccjm de Klerk, J. E., Richards, V., Southgate, A. H., and Blasi, A. (2020). The developmental trajectory of fronto-temporoparietal connectivity as a proxy of the default mode network: a longitudinal fNIRS investigation. *Hum. Brain Mapp.* 41, 2717–2740. doi: 10.1002/hbm.24974

Cava, R., and Orlin, Y. (2022). Home-based transcutaneous tibial nerve stimulation for overactive bladder syndrome: a randomized, controlled study. *Int. Urol. Nephrol.* 54, 1825–1835. doi: 10.1007/s11255-022-03235-z

Chancellor, M. B., Levanovich, P., Rajaganapathy, B. R., and Vereecke, A. J. (2014). Optimum management of overactive bladder: medication vs Botox® vs InterStim® vs urgent® PC. *Urology Practice* 1, 7–12. doi: 10.1016/j.urpr.2014.02.004

Choudhary, M., van Mastrigt, R., and van Asselt, E. (2016). Inhibitory effects of tibial nerve stimulation on bladder neurophysiology in rats. *Springerplus* 5:35. doi: 10.1186/s40064-016-1687-6

Coyne, K. S., Sexton, C. C., Vats, V., Thompson, C., Kopp, Z. S., and Milsom, I. (2011). National community prevalence of overactive bladder in the United States stratified by sex and age. *Urology* 77, 1081–1087. doi: 10.1016/j.urology.2010.08.039

Cui, X., Bray, S., Bryant, D. M., Glover, G. H., and Reiss, A. L. (2011). A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *NeuroImage* 54, 2808–2821. doi: 10.1016/j.neuroimage.2010.10.069

de Groat, W. C. (1998). Anatomy of the central neural pathways controlling the lower urinary tract. *Eur. Urol.* 34, 2–5. doi: 10.1159/000052265

de Groat, W. C., Griffiths, D., and Yoshimura, N. (2015). Neural control of the lower urinary tract. *Compr. Physiol.* 5, 327–396. doi: 10.1002/cphy.c130056

Duan, L., Zhang, Y. J., and Zhu, C. Z. (2012). Quantitative comparison of resting-state functional connectivity derived from fNIRS and fMRI: a simultaneous recording study. *NeuroImage* 60, 2008–2018. doi: 10.1016/j.neuroimage.2012.02.014

Finazzi-Agro, E., Rocchi, C., Pachatz, C., Petta, F., Spera, E., Mori, F., et al. (2009). Percutaneous tibial nerve stimulation produces effects on brain activity: study on the modifications of the long latency somatosensory evoked potentials. *Neurourol. Urodyn.* 28, 320–324. doi: 10.1002/nau.20651

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., and Medvedev, A. V. (2019). Temporal derivative distribution repair (TDDR): a motion correction method for fNIRS. *NeuroImage* 184, 171–179. doi: 10.1016/j.neuroimage.2018.09.025

Fowler, C. J., and Griffiths, D. J. (2010). A decade of functional brain imaging applied to bladder control. *Neurourol. Urodyn.* 29, 49–55. doi: 10.1002/nau.20740

Fu, G., Mondloch, C. J., Ding, X. P., Short, A., Sun, L., and Lee, K. (2014). The neural correlates of the face attractiveness aftereffect: a functional near-infrared spectroscopy (fNIRS) study. *NeuroImage* 85, 363–371. doi: 10.1016/j.neuroimage.2013.04.092

Fukuyama, H., Matsuzaki, S., Ouchi, Y., Yamauchi, H., Nagahama, Y., Kimura, J., et al. (1996). Neural control of micturition in man examined with single photon emission computed tomography using 99mTc-HMPAO. *Neuroreport* 7, 3009–3012. doi: 10.1097/00001756-199611250-00042

Geng, S., Liu, X., Biswal, B. B., and Niu, H. (2017). Effect of resting-state fNIRS scanning duration on functional brain connectivity and graph theory metrics of brain network. *Front. Neurosci.* 11:392. doi: 10.3389/fnins.2017.00392

Griffiths, D. (2015). Neural control of micturition in humans: a working model. *Nat. Rev. Urol.* 12, 695–705. doi: 10.1038/nrurol.2015.266

Griffiths, D., Derbyshire, S., Stenger, A., and Resnick, N. (2005). Brain control of normal and overactive bladder. *J. Urol.* 174, 1862–1867. doi: 10.1097/01.ju.0000177450.34451.97

Griffiths, D., Tadic, S. D., Schaefer, W., and Resnick, N. M. (2007). Cerebral control of the bladder in normal and urge-incontinent women. *NeuroImage* 37, 1–7. doi: 10.1016/j.neuroimage.2007.04.061

Haylen, B. T., de Ridder, D., Freeman, R. M., Swift, S. E., Berghmans, B., Lee, J., et al. (2010). An international Urogynecological association (IUGA)/international continence society (ICS) joint report on the terminology for female pelvic floor dysfunction. *Int. Urogynecol. J.* 21, 5–26. doi: 10.1007/s00192-009-0976-9

Hou, X., Zhang, Z., Zhao, C., Duan, L., Gong, Y., Li, Z., et al. (2021). NIRS-KIT: a MATLAB toolbox for both resting-state and task fNIRS data analysis. *Neurophotonics* 8:010802. doi: 10.1117/1.NPh.8.1.010802

Jiang, Y., Li, Z., Zhao, Y., Xiao, X., Zhang, W., Sun, P., et al. (2020). Targeting brain functions from the scalp: transcranial brain atlas based on large-scale fMRI data synthesis. *NeuroImage* 210:116550. doi: 10.1016/j.neuroimage.2020.116550

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science* 198, 1264–1267. doi: 10.1126/science.929199

Kitta, T., Mitsui, T., Kanno, Y., Chiba, H., Moriya, K., and Shinohara, N. (2015). Brain-bladder control network: the unsolved 21st century urological mystery. *Int. J. Urol.* 22, 342–348. doi: 10.1111/iju.12721

Komesu, Y. M., Ketai, L. H., Mayer, A. R., Teshiba, T. M., and Rogers, R. G. (2011). Functional MRI of the brain in women with overactive bladder: brain activation during urinary urgency. *Female Pelvic Med. Reconstr. Surg.* 17, 50–54. doi: 10.1097/SPV.0b013e3182065507

Matsumoto, S., Ishikawa, A., Matsumoto, S., and Homma, Y. (2011). Brain response provoked by different bladder volumes: a near infrared spectroscopy study. *Neurourol. Urodyn.* 30, 529–535. doi: 10.1002/nau.21016

Matsuta, Y., Mally, A. D., Zhang, F., Shen, B., Wang, J., Roppolo, J. R., et al. (2013). Contribution of opioid and metabotropic glutamate receptor mechanisms to inhibition of bladder overactivity by tibial nerve stimulation. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 305, R126–R133. doi: 10.1152/ajpregu.00572.2012

Nardos, R., Gregory, W. T., Krisky, C., Newell, A., Nardos, B., Schlaggar, B., et al. (2014). Examining mechanisms of brain control of bladder function with resting state functional connectivity MRI. *Neurourol. Urodyn.* 33, 493–501. doi: 10.1002/nau.22458

Nour, S., Svarer, C., Kristensen, J. K., Paulson, O. B., and Law, I. (2000). Cerebral activation during micturition in normal men. *Brain* 123, 781–789. doi: 10.1093/brain/123.4.781

Okada, E., and Delpy, D. T. (2003). Near-infrared light propagation in an adult head model. I. Modeling of low-level scattering in the cerebrospinal fluid layer. *Appl. Opt.* 42, 2906–2914. doi: 10.1364/AO.42.002906

Pang, D., Liao, L., Chen, G., and Wang, Y. (2022). Sacral Neuromodulation improves abnormal prefrontal brain activity in patients with overactive bladder: a possible central mechanism. *J. Urol.* 207, 1256–1267. doi: 10.1097/JU.0000000000002445

Racz, F. S., Stylianou, O., Mukli, P., and Eke, A. (2018). Multifractal dynamic functional connectivity in the resting-state brain. *Front. Physiol.* 9:1704. doi: 10.3389/fphys.2018.01704

Reynolds, W. S., Fowke, J., and Dmochowski, R. (2016). The burden of overactive bladder on US public health. *Curr. Bladd. Dysfunct. Rep.* 11, 8–13. doi: 10.1007/s11884-016-0344-9

Sakakibara, R., Hattori, T., Yasuda, K., and Yamanishi, T. (1996). Micturitional disturbance after acute hemispheric stroke: analysis of the lesion site by CT and MRI. *J. Neurol. Sci.* 137, 47–56. doi: 10.1016/0022-510X(95)00322-S

Sato, T., Ito, M., Suto, T., Kameyama, M., Suda, M., Yamagishi, Y., et al. (2007). Time courses of brain activation and their implications for function: a multichannel near-infrared spectroscopy study during finger tapping. *Neurosci. Res.* 58, 297–304. doi: 10.1016/j.neures.2007.03.014

Schneider, M. P., Gross, T., Bachmann, L. M., Blok, B. F., Castro-Diaz, D., del Popolo, G., et al. (2015). Tibial nerve stimulation for treating neurogenic lower urinary tract dysfunction: a systematic review. *Eur. Urol.* 68, 859–867. doi: 10.1016/j.eururo.2015.07.001

Stewart, W. F., van Rooyen, J., Cundiff, G. W., Abrams, P., Herzog, A. R., Corey, R., et al. (2003). Prevalence and burden of overactive bladder in the United States. *World J. Urol.* 20, 327–336. doi: 10.1007/s00345-002-0301-4

Tadic, S. D., Griffiths, D., Schaefer, W., Murrin, A., Clarkson, B., and Resnick, N. M. (2012). Brain activity underlying impaired continence control in older women with overactive bladder. *Neurourol. Urodyn.* 31, 652–658. doi: 10.1002/nau.21240

Te Dorsthorst, M., van Balken, M., and Heesakkers, J. (2020). Tibial nerve stimulation in the treatment of overactive bladder syndrome: technical features of latest applications. *Curr. Opin. Urol.* 30, 513–518. doi: 10.1097/MOU.0000000000000781

Xu, S. Y., Lu, F. M., Wang, M. Y., Hu, Z. S., Zhang, J., Chen, Z. Y., et al. (2020). Altered functional connectivity in the motor and prefrontal cortex for children with Down's syndrome: an fNIRS study. *Front. Hum. Neurosci.* 14:6. doi: 10.3389/fnhum.2020.00006

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 04 May 2023
DOI 10.3389/fnins.2023.1183145

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Lei Xie,
Zhejiang University of Technology, China
Lu Zhang,
University of Texas at Arlington, United States

*CORRESPONDENCE
Bao Ge
✉ bob_ge@snnu.edu.cn

RECEIVED 09 March 2023
ACCEPTED 17 April 2023
PUBLISHED 04 May 2023

CITATION
He M, Hou X, Ge E, Wang Z, Kang Z, Qiang N, Zhang X and Ge B (2023) Multi-head attention-based masked sequence model for mapping functional brain networks. *Front. Neurosci.* 17:1183145. doi: 10.3389/fnins.2023.1183145

Multi-head attention-based masked sequence model for mapping functional brain networks

Mengshen He[1,2], Xiangyu Hou[2], Enjie Ge[2], Zhenwei Wang[2], Zili Kang[2], Ning Qiang[2], Xin Zhang[3] and Bao Ge[1,2]*

[1]Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, China, [2]School of Physics and Information Technology, Shaanxi Normal University, Xi'an, China, [3]Institute of Medical Research, Northwestern Polytechnical University, Xi'an, Shaanxi, China

The investigation of functional brain networks (FBNs) using task-based functional magnetic resonance imaging (tfMRI) has gained significant attention in the field of neuroimaging. Despite the availability of several methods for constructing FBNs, including traditional methods like GLM and deep learning methods such as spatiotemporal self-attention mechanism (STAAE), these methods have design and training limitations. Specifically, they do not consider the intrinsic characteristics of fMRI data, such as the possibility that the same signal value at different time points could represent different brain states and meanings. Furthermore, they overlook prior knowledge, such as task designs, during training. This study aims to overcome these limitations and develop a more efficient model by drawing inspiration from techniques in the field of natural language processing (NLP). The proposed model, called the Multi-head Attention-based Masked Sequence Model (MAMSM), uses a multi-headed attention mechanism and mask training approach to learn different states corresponding to the same voxel values. Additionally, it combines cosine similarity and task design curves to construct a novel loss function. The MAMSM was applied to seven task state datasets from the Human Connectome Project (HCP) tfMRI dataset. Experimental results showed that the features acquired by the MAMSM model exhibit a Pearson correlation coefficient with the task design curves above 0.95 on average. Moreover, the model can extract more meaningful networks beyond the known task-related brain networks. The experimental results demonstrated that MAMSM has great potential in advancing the understanding of functional brain networks.

KEYWORDS

masked sequence modeling, multi-head attention, functional brain networks, feature selection, task fMRI

1. Introduction

Research into the function of the human brain has garnered significant attention and has been a popular field of study for several decades. One pivotal research direction in this field is the mapping of functional brain networks (FBNs), which has become a useful way to study the working mechanisms of the brain. By providing insight into the underlying neural mechanisms of such networks, FBNs hold the potential to unravel the working of the brain

(Power et al., 2010; Park and Friston, 2013; Sporns and Betzel, 2016; Jiang et al., 2021), as well as the pathogenesis of several diseases (Canario et al., 2021). Therefore, exploring FBNs is crucial for comprehending the complex dynamics of the brain and can offer an avenue for further understanding the neural processes underlying different functions.

In traditional methods, generalized linear models (GLM) (Beckmann et al., 2003; Barch et al., 2013), independent component analysis (ICA) (McKeown, 2000; Beckmann et al., 2005; Calhoun and Adali, 2012), and sparse dictionary learning (SDL) (Lv et al., 2014; Ge et al., 2016; Lee et al., 2016; Zhang et al., 2016; Shen et al., 2017; Zhang et al., 2018) have been utilized to construct functional brain networks. Moreover, other machine learning techniques have been effectively applied to fMRI data analysis, such as support vector machines (SVM) (LaConte et al., 2005; Mourao-Miranda et al., 2006) for fMRI analysis and classification, and principal component analysis (PCA) (Thirion and Faugeras, 2003; Smith et al., 2014) for fMRI data dimensionality reduction. With the advancement of deep learning technology, numerous deep learning models have been applied to fMRI data analysis and functional brain network construction. For instance, Huang et al. (2017) proposed a deep convolutional autoencoder (DCAE) to extract hierarchical features from fMRI data; Zhao et al. (2018) proposed a spatiotemporal convolutional neural network (ST-CNN) to learn temporal and spatial information from fMRI data simultaneously; Qiang et al. (2020) proposed a spatiotemporal self-attention mechanism (STAAE) (Dong et al., 2020b) for brain functional network modeling and ADHD disease classification. Additionally, Qiang et al. (2020) proposed a residual autoencoder (RESAE) (Dong et al., 2020a) for constructing task related functional brain networks. Jiang et al. (2023) introduce a Spatio-Temporal Attention 4D Convolutional Neural Network (STA-4DCNN) model to characterize individualized spatio-temporal patterns of FBNs. Yan et al. (2022) proposed a Multi-Head Guided Attention Graph Neural Network (Multi-Head GAGNN) to simultaneously model both spatial and temporal patterns of holistic functional brain networks. Experimental results have indicated that deep learning methods are effective in fMRI data modeling and brain network construction tasks, which demonstrate the significant advantages of deep learning models.

Although the methods mentioned above have shown promising results, there are still certain limitations that need to be addressed. Firstly, the current design and parameterization of models do not fully account for the characteristics of fMRI data. For instance, the same signal value at different time points may have different meanings depending on the task or state, and thus, it is crucial to exploit this information for improving model performance. Secondly, the model training process disregards some prior knowledge, such as task design curves, which could potentially enhance the efficacy and efficiency of the model. These limitations underscore the need for more advanced techniques that can tackle these challenges and improve the accuracy and applicability of fMRI analysis.

Recent research has revealed the exceptional capabilities of Transformer models (Vaswani et al., 2017) in tasks such as text analysis and prediction. One of key mechanisms of transformer is to use multi-head attention to do the processing of sequence data. By leveraging multi-head attention mechanisms, the distinctive semantics of a single word in different language contexts can be analyzed. For instance, the term "apple" could signify either a fruit or a mobile phone brand in various language contexts. Given the similarity between fMRI time series and text sequences, multi-head attention mechanisms can be employed to extract features from fMRI data. Furthermore, the growing popularity of the masked language modeling (MLM) training method in the Bert model (Devlin et al., 2018) suggests that masking-based training techniques are remarkably effective at capturing contextual information. Since there are similarities between fMRI time series and sentences, the multi-head attention mechanism and mask training method can be extended to fMRI feature extraction.

So, this manuscript proposed a novel model called the Multi-head Attention-based Masked Sequence Model (MAMSM) which utilizes a multi-head attention mechanism to scrutinize different states of voxel signals at various locations while also implementing the Masked Sequence Model (MSM) method to analyze and process the fMRI time series. Furthermore, MAMSM employs both randomly discrete and continuous masks in the masking operation to enhance the model's learning capacity and training effectiveness. In addition to that, this study leverages prior knowledge of the task design curves and cosine similarity to construct a new loss function, resulting in improved outcomes in model training.

In order to demonstrate the effectiveness of our proposed model, we utilized data from the Human Connectome Project (HCP) (Van Essen et al., 2013) and analyzed the seven task-state datasets of 10 individuals using both individual and group average approaches. To evaluate the performance of our model, we compared it with the SDL and STAAE methods. The experimental results indicate that the FBNs extracted by our proposed model outperformed those extracted by the other methods across various task datasets. Notably, our model also detected several brain networks that were distinct from the task-state-corresponding FBNs, and we subsequently identified some networks as similar to the known resting-state brain networks. Specifically, our experimental results demonstrate that our model is highly effective in extracting features from a small amount of data, which is particularly important in the context of brain imaging research where data acquisition is often difficult, costly, and resource-intensive. A brief version of the study has been published as a conference paper in the MICCAI 2022 (He M. et al., 2022).

2. Materials and methods

2.1. Overview

As shown in **Figure 1**, the proposed method consists of three main steps: (1) four-dimensional fMRI data is pre-processed and mapped to two-dimensional space; (2) the pre-processed two-dimensional fMRI time series is input into the MAMSM, composed of multiple headed attention mechanisms, and trained with a mask-based approach; (3) all the latent features extracted from the pre-training are input into the feature selection layer, which are trained with a loss function by leveraging the prior task designs. Finally, the features output by the encoder of the feature selection layer are regressed by lasso and mapped back to the original brain space, resulting in the visualization of FBNs.

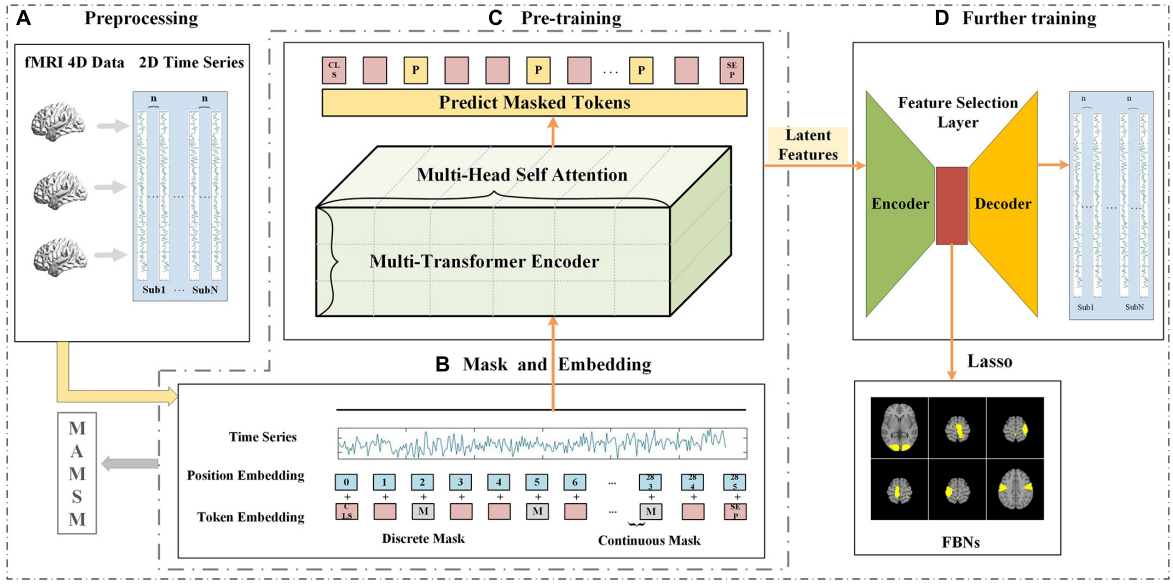


FIGURE 1
The framework of MAMSM **(A)** preprocessing, which involves mapping all subjects' tfMRI data to 2D space; **(B)** masking and embedding operations of time-series data; **(C)** pre-training of the model, consisting of three layers of transformer encoders; **(D)** further training is performed using the feature selection layer, and the obtained features are used to map FBNs.

TABLE 1 Summary of used datasets.

	H	W	D	Time points	Voxels	Training subjects
Motor	46	55	46	284	28,546	10
Emotion	46	55	46	176	28,546	10
Gambling	46	55	46	253	28,546	10
Language	46	55	46	316	28,546	10
Relational	46	55	46	232	28,546	10
Social	46	55	46	274	28,546	10
WM	46	55	46	405	28,546	10

2.2. Materials and pre-processing

The dataset from the Human Connectome Project Q3 was used in this work, which is publicly available on the website.[1] We selected randomly the 10 subjects from HCP dataset. To evaluate the temporal features and spatial features obtained by the MAMSM, we chose 24 task designs from seven tasks. The corresponding hemodynamic response function (HRF) responses, which are the convolution of the task paradigm and HRF function, are utilized as temporal templates and the group-wise functional brain networks (FBNs) derived from the GLM are utilized as spatial templates (Güçlü and Van Gerven, 2017). For the sake of description, 24 distinct symbols were used to represent each of the selected task designs. For emotion task, E1 is for emotional faces, and E2 is for simple shapes. For gambling task, G1 is for punishment over baseline, and G2 is for reward over baseline. For language task, L1 is for math over story, and L2 is for story over math. For social task, S1 is for social over baseline, and S2 is for random over baseline. For relational task, R1 is for match over baseline, and R2 is for relational over baseline. For motor task, M1-M6 are for cue, left foot movement, left hand movement, right foot movement, right hand movement, and tongue movement, respectively. For working memory task, W1-W8 are for the 2-back and 0-back task events of body parts, places, faces, and tools, respectively.

The parameters of data collection used in this text is as follows: a 90 × 104 matrix, 220 mm FOV, 72 slices, TR = 0.72 s, TE = 33.1 ms, Flip angle = 52°, BW = 2,290 Hz/Px, in-plane FOV = 208 mm × 180 mm. For the tfMRI data, the pre-processing operations included skull stripping, motion correction, slice timing correction, spatial smoothing, global drift removal (high pass filtering) and registration to MNI space. Table 1 provides an overview of the pre-processed task functional magnetic resonance imaging (tfMRI) datasets used in this study. After pre-processing of the tfMRI data, the four-dimensional tfMRI data was transformed into a two-dimensional matrix by using Nilearn tools (available at https://nilearn.github.io/) and the MNI-152 mask. Data for each time point comprised 28,546 voxels.

1 https://db.humanconnectome.org

TABLE 2 The results of training with different mask operations.

Mask strategies	Training loss	Predict loss
Discrete	0.043	4.73
Continuous	0.047	4.739
Discrete and continuous	0.04	4.719

2.3. MAMSM

2.3.1. MSM

In recent years, Masked Language Modeling (MLM) and Masked Image Modeling (MIM) approaches have been widely employed in Natural Language Processing (NLP) (Devlin et al., 2018; Chung et al., 2021; Sinha et al., 2021) and Computer Vision (CV) (Zhou et al., 2021; He K. et al., 2022; Tong et al., 2022; Xie et al., 2022) due to their demonstrated efficacy in extracting contextual information through mask training. This work utilized Masked Sequence Modeling (MSM) to process fMRI sequence data. MSM is a self-supervised training method in which a portion of the tokens in the sequence are replaced with [mask] symbols and the remaining tokens and location information are used to predict the tokens replaced with [mask]. This training method allows the model to learn more about the relationships between contexts.

In the BERT model proposed by Devlin et al. (2018), the [CLS] (Classification Token) serves to create a compact representation of the entire input sequence. This condensed representation can be used for tasks such as text classification and similarity computation. Specifically, for each input fMRI time series, the proposed model is designed to generate a vector representation for each input. By adding the special [CLS] tag at the beginning of the sequence, this vector representation of the tag serves as a summary of the entire sequence, compressing and integrating the information from the entire input. As a result, the [CLS] tag provides a comprehensive representation for subsequent feature extraction and similarity calculation processes.

Before the mask processing process, the fMRI data was normalized to a range of (0, 1). After normalization, we retained three decimal places for the values, resulting in a maximum of 1,001 distinct values (from 0, 0.001, 0.002, . . . to 1) for the whole-brain signals. In the subsequent model training process, we treat these 1,001 different values as 1,001 classes, simplifying the model training process into a multi-classification problem. That is, if we want to predict the value of fMRI signals at a certain time point, we converted it into categories with a total of 1,001 values for classification. The prediction range of the model is also within these 1,001 classes of values. When predicting the value of a masked position, the model only needs to determine the class to which it belongs. To facilitate the prediction of token values, a multi-classification task was employed, where in a cross-entropy loss function was utilized to compute the error between the model's predicted value and the actual value. As shown below, where $\mathbf{y_i}$ is the true probability distribution, $\widehat{\mathbf{y}}_\mathbf{i}$ is the predicted probability distribution, and n is the number of categories:

$$\mathbf{C}E\left(\mathbf{y_i}, \widehat{\mathbf{y}}_\mathbf{i}\right) = -\sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} \mathbf{y_i} log\left(\widehat{\mathbf{y}}_\mathbf{i}\right)$$

In the mask processing process, for each fMRI sequence input, a certain proportion of positions on the fMRI time series will be randomly covered, with the original signal values replaced by [mask]. Here, taking the tfMRI sequence of Motor task as an example, we selected roughly 10% time points as mask locations for each input with the length of 284 time steps, as illustrated in Figure 1B. After the Mask operation was performed, the pre-training stage in the proposed model employed an unsupervised training process to predict the token values of the masked locations, as shown in Figure 1C.

In order to enhance the learning capability of the model and achieve optimal training outcomes, this study employs a combination of continuous and discrete masking techniques. When using only discrete masking, the model may be able to predict the values of the masked regions through simple methods such as averaging the values of its previous and subsequent time steps. This may lead to the model failing to learn deeper features. To avoid this issue, we designed more sophisticated methods of masking, such as continuous mask, etc., Table 2 presents the outcomes of the training with different masking modes, where 90% of the voxels in the same subject are allocated for the training set, 10% for the test set, and the same training parameters are utilized. We adopt a uniform sampling strategy for voxel selection, wherein every ten voxels, the first nine are assigned to the training set, and the last one is designated for the testing set. By comparing the

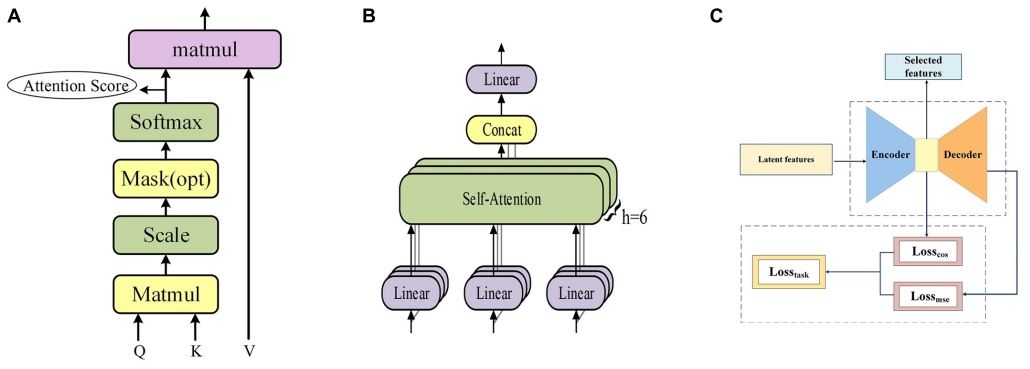


FIGURE 2
(A) The frame of self-attention. **(B)** The frame of multi-head attention. **(C)** The frame of feature selection layer.

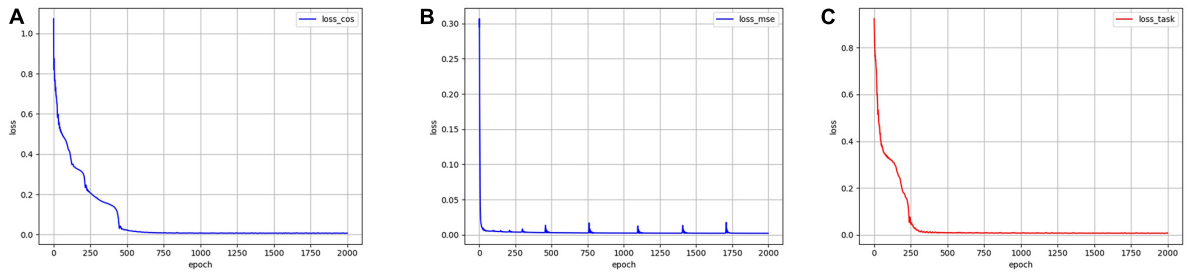


FIGURE 3
The training errors by using three different loss functions. **(A)** Error variation by using $Loss_{cos}$. **(B)** Error variation by using $Loss_{mse}$. **(C)** Error variation by using $Loss_{task}$.

minimum loss on the training set and the test set, it can be seen that the combination of the two mask operations can achieve better results.

2.3.2. Multi transformer encoder layers

The Transformer model is a sophisticated deep neural network that is based on an attention mechanism, originally introduced by Vaswani et al. (2017) for machine translation. The model is structured according to the seq2seq paradigm and comprises two primary components: an encoder that encodes the input sequence and a decoder that generates the output sequence. Unlike traditional Recurrent Neural Network (RNN) models (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997; Graves et al., 2005; Cho et al., 2014), the transformer model utilizes multi-head attention mechanism for computation. This mechanism can represent information from multiple semantic spaces, capturing different meanings of the same words in different contexts, similar to the same signal values in fMRI data may represent different states and meanings.

Therefore, in this manuscript, each fMRI sequence is embedded and masked as the input of the transformer encoder, and then the input is linearly transformed to obtain three matrices, namely **Q** (Query), **K** (Key) and **V** (Value). Subsequently, **Q** and **K** are dot-multiplied and then normalized by dividing by $\sqrt{\mathbf{d_k}}$ to stabilize the gradient. Subsequently, a softmax operation was used to obtain the attention score, which represents the importance of each position of the fMRI sequence, and then multiplied by V to obtain the output of self-attention, as shown in Figure 2A. Eventually, the output of multiple self-attentions is superimposed as the output of multi-headed attention, as shown in Figure 2B. The formulae of self-attention and multi-head attention can be expressed as follows, where $\mathbf{h}ead_{\mathbf{i}}$ denotes the $\boldsymbol{i}$-th self-attention mechanism.

$$\mathbf{M}ultiHead\,(\mathbf{Q},\mathbf{K},\mathbf{V}) \;=\; \mathbf{C}oncat\left(\mathbf{h}ead_{\mathbf{1}},,\mathbf{h}ead_{\mathbf{n}}\right)\mathbf{W^O}$$

$$\mathbf{h}ead_{\mathbf{i}} \;=\; \mathbf{A}ttention(\mathbf{Q},\mathbf{K},\mathbf{V})$$

$$\mathbf{A}ttention\,(\mathbf{Q},\mathbf{K},\mathbf{V}) \;=\; \mathbf{s}oftmax(\frac{\mathbf{QK^T}}{\sqrt{\mathbf{d_k}}})\mathbf{V}$$

Upon completion of the pre-training of the model, the attention score was extracted as a feature matrix, which represents the weights at various time points within an fMRI time series. After the model pre-training was completed, the attention scores were extracted as the features representing the weights of each time point in the fMRI time series. We use the sliding average operation to smooth the attention scores, and then use the average results as latent features of the pre-trained model. We set the size of the sliding average window to 10 and the step size of the sliding window is 1.

2.3.3. Feature selection layer

Here we propose a novel loss function, $\boldsymbol{Loss_{task}}$, for the training of a feature selection layer in autoencoders, as illustrated in Figure 2C. By combining mean squared loss function ($\boldsymbol{Loss_{mse}}$) and cosine similarity loss function ($\boldsymbol{Loss_{cos}}$), this loss function is more conducive to the task of tfMRI data compared to the other methods (Dong et al., 2020b; Qiang et al., 2020), which often focus solely on reconstruction error such as MSE, disregarding the latent feature distribution and the relationship with the task curves, both of which are indispensable to characterize fMRI time series. The latent feature matrix obtained from pre-training serves as the input for the encoder, which, after training, produces the final feature matrix as its output. Through this process, the feature selection layer also facilitates the reduction of dimensionality of the latent feature matrix, thus contributing to more efficient and effective features. The $\boldsymbol{Loss_{task}}$ function is formulated as the combination of $\boldsymbol{Loss_{cos}}$ and $\boldsymbol{Loss_{mse}}$ and we experimentally chose the value of k to 1 in this work, as follows:

$$\boldsymbol{Loss}_{\mathbf{mse}} \;=\; \mathbf{M}SE \;=\; \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i\,=\,1}}^{\mathbf{m}}\mathbf{w_i}(\mathbf{y_i}-\widehat{\mathbf{y}}_\mathbf{i})^\mathbf{2}$$

$$\boldsymbol{Loss}_{\mathbf{cos}} \;=\; \mathbf{l_i} \;=\; \mathbf{1}-\mathbf{c}os\left(\mathbf{x_i},\mathbf{y_i}\right)$$

$$\boldsymbol{Loss}_{\mathbf{task}} \;=\; \mathbf{l}oss\,(\mathbf{mse})+\mathbf{k}*\mathbf{l}oss(\mathbf{c}os)$$

The actual value $\mathbf{y_i}$ and the predicted value $\widehat{\mathbf{y}}_\mathbf{i}$ are compared by calculating the cosine similarity between the n sequences of the

TABLE 3 The final training errors of three different loss functions.

	$Loss_{cos}$	$Loss_{mse}$	$Loss_{task}$
Cos-error	0.0074	1.0632	**0.0067**
Mse-error	0.2955	0.0022	**0.0022**

The bold values represent the minimum values of each row.

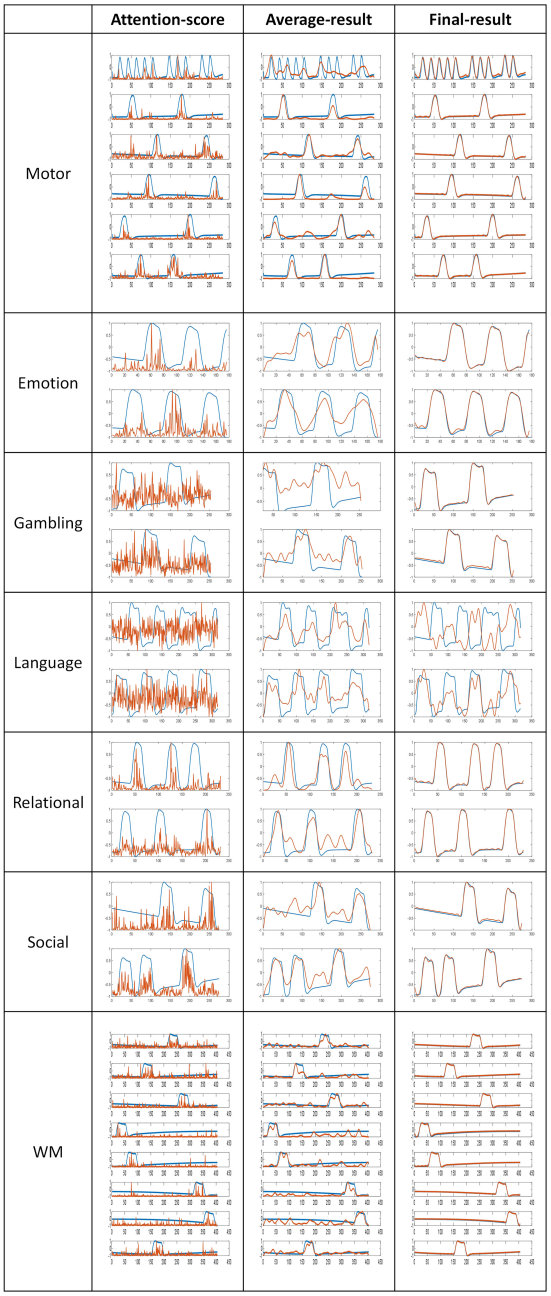


FIGURE 4
Comparison of features and task designs. The blue curves represent the task designs, the red curves depict the features.

feature$\mathbf{x_i}$ and the n task design curves $\mathbf{y_i}$, and the cosine similarity calculation formula is:

$$cos\left(\mathbf{x_i}, \mathbf{y_i}\right) = \frac{\sum_{i=1}^{n}\left(\mathbf{x_i} \times \mathbf{y_i}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathbf{x_i}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(\mathbf{y_i}\right)^2}}$$

Mse-error is the MSE reconstruction error of the decoder output and the original data; Cos-error is the cosine similarity error between the n sequences of the encoder output feature and n task design curves. To demonstrate the effectiveness of the proposed new loss function, an ablation experiment was conducted using the same data and parameters. The model was trained using $\boldsymbol{Loss_{cos}}$, $\boldsymbol{Loss_{task}}$, and $\boldsymbol{Loss_{mse}}$, respectively. As illustrated in Figure 3, when $\boldsymbol{Loss_{task}}$ was used, the convergence rate was faster and more stable than when only $\boldsymbol{Loss_{cos}}$ or $\boldsymbol{Loss_{mse}}$ was used. Quantitatively, Table 3 shows that when $\boldsymbol{Loss_{task}}$ was employed, the final Cos-error and Mse-error were lower.

2.3.4. Mapping FBNs

To obtain the spatial distribution of the functional network, lasso regression is applied to the feature matrix and the original two-dimensional input data to get the sparse coefficient matrix, which represents the spatial distribution of the functional network The calculation formula of LASSO regression (Pedregosa et al., 2011) is as follows:

$$\min_{\mathbf{w}} \frac{1}{2\mathbf{T}} \| \mathbf{Z} - \mathbf{X}W \|_2^2 + \lambda \| \mathbf{W} \|_1$$

$\mathbf{Z}$ is the original 2D input data, $\boldsymbol{T}$ represents the total number of time points, $\boldsymbol{X}$ is the feature matrix, and $\boldsymbol{W}$ is the regressed sparse coefficient matrix. The coefficient matrix $\boldsymbol{W}$, which captures the spatial distribution information of the underlying functional network, was then mapped back to the original 3D brain image space, the result was finally visualized as FBNs.

3. Results

The work reports its findings in terms of two primary dimensions: temporal and spatial features. To evaluate temporal features, the final feature matrix was utilized to obtain partial task-related features, which were subsequently evaluated for similarity with the task design curves. Spatial features were assessed by computing the similarity between the derived FBNs and the templates derived from the GLM. Besides task-related FBNs, we also identified additional FBNs, including those resting-state FBNs.

TABLE 4 Pearson correlation coefficient between the features and the task designs.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	/
Attention-score	0.331	0.343	0.321	0.333	0.287	0.268	0.270	0.267	0.276	0.262	0.275	0.276	/
Average-result	0.851	0.894	0.792	0.792	0.813	0.799	0.870	0.804	0.845	0.845	0.789	0.856	/
Final-result	0.999	0.998	0.998	0.998	0.999	0.999	0.999	0.994	0.999	0.999	0.998	0.999	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
Attention-score	0.262	0.250	0.319	0.583	0.419	0.337	0.336	0.423	0.430	0.408	0.427	0.567	0.345
Average-result	0.737	0.785	0.796	0.916	0.880	0.889	0.451	0.903	0.913	0.880	0.873	0.961	0.831
Final-result	0.727	0.737	0.998	0.997	0.999	0.999	0.973	0.997	0.998	0.997	0.997	0.997	0.975

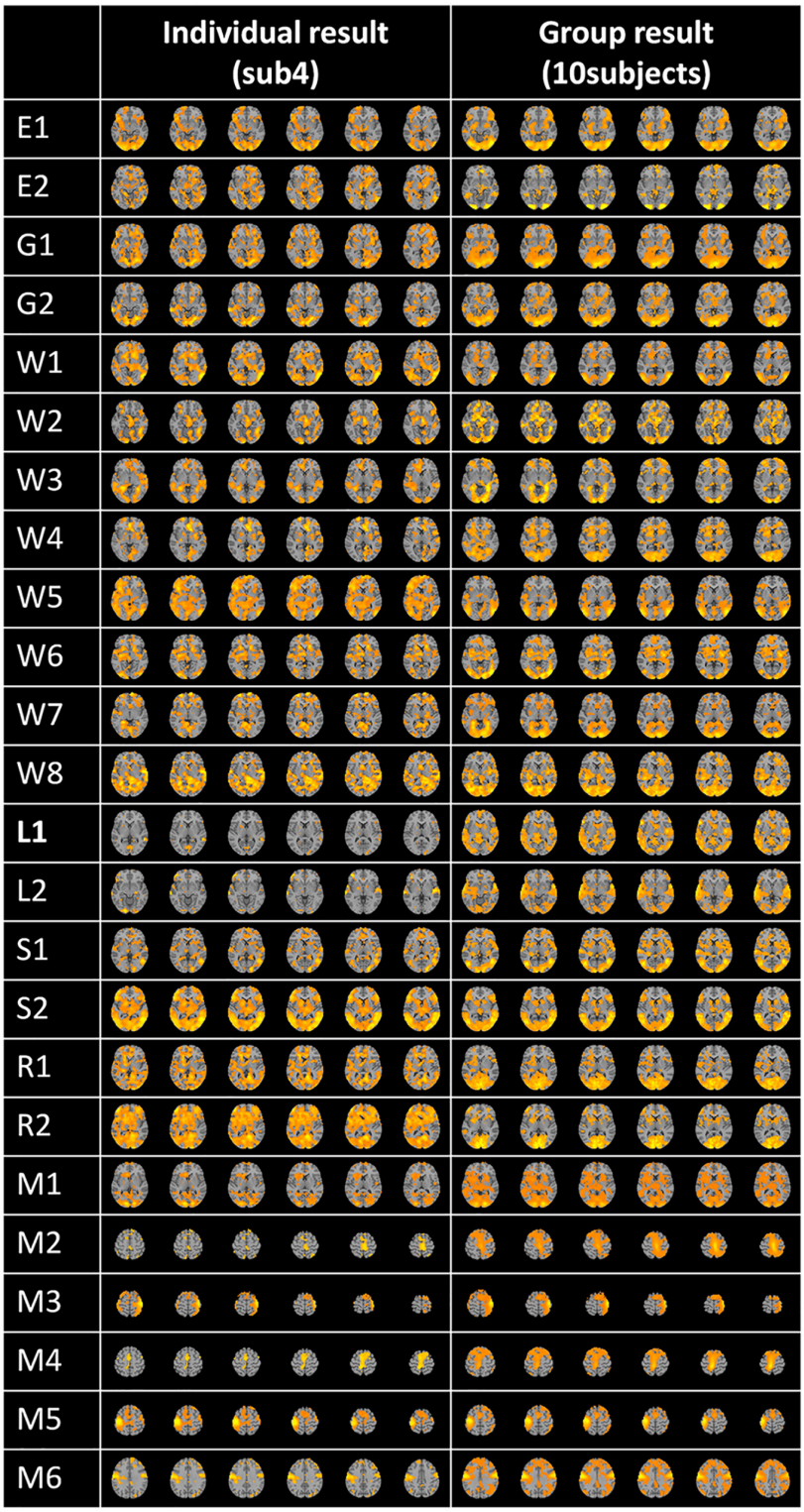


FIGURE 5
Individual and group averaged FBNs.

3.1. Temporal features

The proposed model generated three different temporal feature matrices, namely the intermediate "attention-score" feature, which is obtained immediately after model pre-training; the "average-result" feature, calculated by computing a sliding average of the attention-score feature; and the "Final-result" feature, obtained after training the feature selection layer. The dimension of attention-score, average-result, and final-result are [6*28,546,t], [6*28,546,t], and [256,t]. In this work, "t"

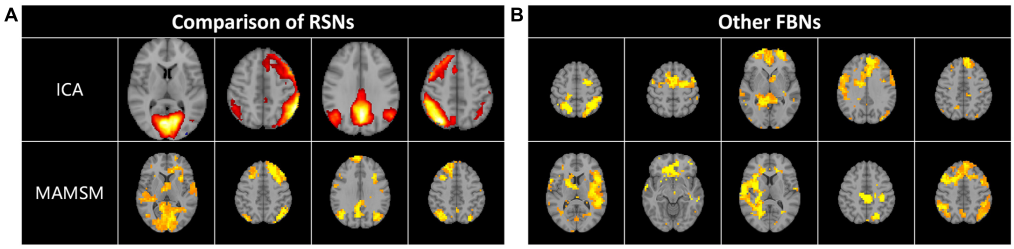


FIGURE 6
(A) Some resting-state FBNs. **(B)** Other FBNs.

represents the length of the fMRI sequence's time dimension corresponding to different tasks, while "6" denotes the number of attention heads we have set for the multi-head attention mechanism. To evaluate the significance of the three kinds of features selected in this study, a comparative analysis is conducted between these features and the task design curves. As illustrated in Figure 4, a graphical representation of the three kinds of features and the correspondingly relevant task design curves are presented. The blue curves represent the task design curves and serve as the baseline, the red curves depict the features.

Based on the results of the comparison, it is evident that the attention-score and task curves display an obvious fitting trend, with their highest peak approximately coinciding with the peak of the task design curves. Furthermore, the application of a sliding average filter results in an even higher similarity between the average-result and task design curves. These outcomes provide evidence that the latent features derived from the pre-training module are both meaningful and interpretable.

In order to quantitatively compare the similarity between the feature matrices and the task design curves, the Pearson correlation coefficient was calculated in this work, the formula for the Pearson correlation coefficient is presented below:

$$\rho(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{n}(\mathbf{X_i}-\mu_X)(\mathbf{Y_i}-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(\mathbf{X_i}-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{Y_i}-\mu_Y)^2}}$$

where $\mathbf{X}$, $\mathbf{Y}$ are the features and task design curves, μ_X, μ_Y are the mean of the them and$\mathbf{X_i}$, $\mathbf{Y_i}$ are the samples of them.

The Pearson correlation coefficient values serve as an indicator of the strength of the correlation, with higher values indicating stronger correlations. As shown in Table 4, all Pearson's correlation coefficients achieved statistical significance at the level of $P < 0.05$. These results demonstrate that the features extracted by the proposed pre-training model were significantly correlated with the design curves. Specifically, the initially extracted attention-score feature exhibited a certain degree of similarity with the task design curves. With the application of the sliding average technique, the Average-result feature approached the task design curves more. Finally, the incorporation of a feature selection layer and a new loss function as a guide led to the generation of the Final-result feature. The Pearson correlation coefficient for the task design curves was significantly improved from 0.831 to 0.975 as a result. These findings underscore the importance of the pre-training model and feature selection layer, and provide further support for the efficacy and interpretability of the proposed model in this study.

3.2. Spatial features

3.2.1. Task FBNs

Following the feature selection process, the feature matrix was remapped to the original 3D brain space for the visualization of FBNs using lasso regression, as shown in Figure 5. This figure displays a randomly selected individual FBN for 24 tasks and group-averaged FBNs from 10 subjects. As demonstrated in Figure 5, each task-related FBN can be accurately identified, and the FBNs becomes even more pronounced after group averaging.

3.2.2. Other FBNs

Multi-head Attention-based Masked Sequence Model can not only acquire the known activated networks, but also enable the identification of other brain networks with specific patterns. In this work, we also selected and displayed a part of them. After comparison and analysis, we found some resting-state networks, which were compared and displayed with the corresponding resting-state brain network templates obtained by the ICA method, as shown in Figure 6A. In addition, this manuscript also displays other brain networks with certain patterns, as shown in Figure 6B.

3.3. Comparative experiments

To further evaluate the effectiveness of the proposed MAMSM, it is compared with SDL (Lv et al., 2015)and STAAE (Dong et al., 2020b). SDL is the traditional way to build FBNs. STAAE has been proposed as a deep learning method recently. All three methods are applied to the same dataset and their temporal and spatial characteristics are compared in this section.

3.3.1. Comparison of temporal features

In this study, three different methods were employed for comparison purposes. In order to ensure fairness in our comparison analysis, we adopted the "average-result" features instead of the "final result" features for comparison with the features obtained from SDL and STAAE, as our proposed model leveraged prior knowledge (task designs) to train the model in the feature selection layer. Figure 7 displays the task design curves, with the blue curves representing specific task design curves used as comparison benchmarks and the red curves representing the task-related features. Our qualitative and quantitative comparison analysis aimed to assess the degree of correlation between these two curves. For quantitative comparison, the Pearson correlation

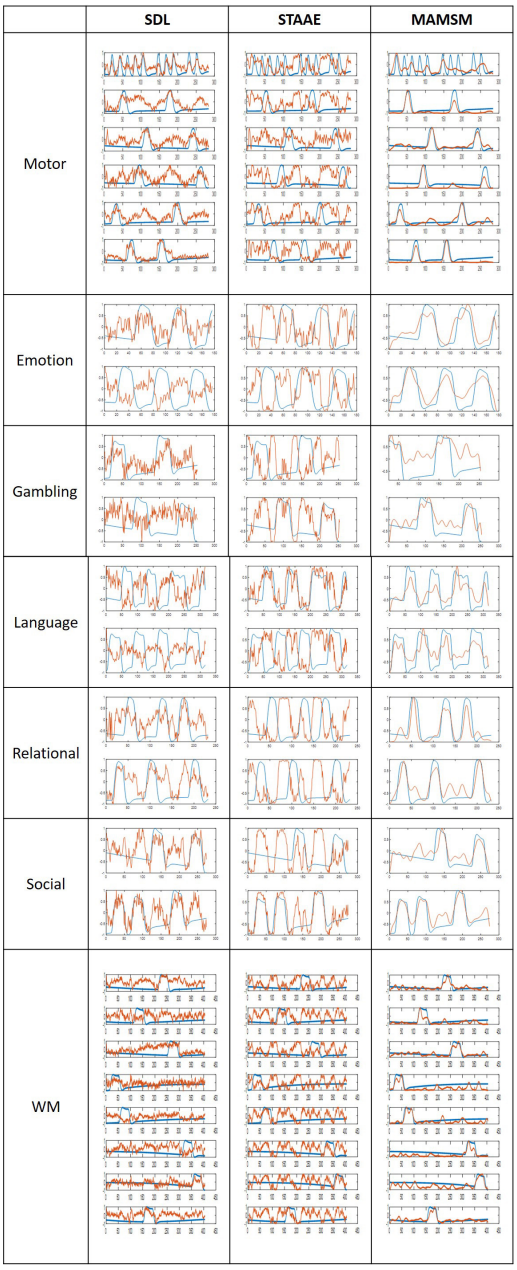


FIGURE 7
Comparison of features and task designs obtained by different methods. The blue curves represent the task designs, the red curves depict the features.

coefficient was employed to assess the similarity between the extracted features and the task curves, as presented in Table 5. It should be noted that Figure 7 shows the results of an individual. Table 5 is the average result of ten individuals.

As shown in Figure 7, the correlation between the features generated by MAMSM and the task design curves was found to be significantly higher compared to that between the features generated by SDL/STAAE and the task design curves. Quantitatively, the results presented in Table 5 demonstrate that the proposed MAMSM achieved a higher averaged Pearson correlation coefficient (0.824) compared to that from SDL (0.527) or STAAE (0.306). Overall, the results of our experiment demonstrate the effectiveness of MAMSM for constructing FBNs based on tfMRI.

In terms of individual-level performance, our results indicate that the deep learning method STAAE performed slightly worse than SDL and MAMSM. It is worth noting that according to the description of the STAAE (Dong et al., 2020b), the method can achieve better results when applied to larger datasets. However, the inherent requirement of deep learning methods for large volumes of data may limit their advantage over traditional methods in cases where data availability is limited. Our proposed method, on the other hand, demonstrates good performance on individual data, suggesting that it can effectively learn temporal features from small datasets.

3.3.2. Comparison of spatial features

In order to qualitatively compare the spatial features from the three methods, this work applies SDL, STAAE, and MAMSM to the same dataset and obtains the group averaged results, as shown in Figure 8. The GLM templates were derived by summarizing a large amount of individual data and were subsequently employed for the purpose of comparing the performance of FBNs generated through various methods. Our results demonstrate that the activation maps obtained through MAMSM exhibit greater resemblance to the GLM templates.

Quantitatively, we also used the spatial overlap rate as an indicator to compare the FBNs from the three methods and the GLM template. The spatial overlap rate can be used to compare the similarity between two different networks, which is defined as follows:

$$\mathbf{OR}(\mathbf{N^1}, \mathbf{N^2}) = \frac{\sum_{\mathbf{i}=1}^{\mathbf{n}} |\mathbf{N_i}^1 \cap \mathbf{N_i}^2|}{\sum_{\mathbf{i}=1}^{\mathbf{n}} |\mathbf{N_i}^1 \cup \mathbf{N_i}^2|}$$

$N1$, $N2$ are the two brain networks to be compared, n is the number of voxel points of the brain network. The spatial overlap

TABLE 5 Pearson correlation coefficient obtained by SDL, STAAE, and MAMSM.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	
SDL	0.631	0.624	0.483	0.515	0.390	0.356	0.395	0.443	0.419	0.369	0.453	0.379	/
STAAE	0.322	0.246	0.351	0.385	0.195	0.128	0.259	0.272	0.088	0.069	0.197	0.155	/
MAMSM	0.830	0.867	0.848	0.821	0.864	0.870	0.869	0.803	0.849	0.819	0.799	0.869	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
SDL	0.603	0.622	0.523	0.673	0.514	0.564	0.658	0.603	0.586	0.493	0.603	0.738	0.527
STAAE	0.606	0.619	0.302	0.651	0.440	0.422	0.429	0.218	0.203	0.189	0.226	0.383	0.306
MAMSM	0.760	0.777	0.812	0.835	0.863	0.868	0.500	0.836	0.862	0.838	0.850	0.875	0.824

TABLE 6 The spatial overlap rate obtained by SDL, STAAE, and MAMSM.

	E1	E2	G1	G2	W1	W2	W3	W4	W5	W6	W7	W8	
SDL	0.150	0.102	0.231	0.200	0.231	0.236	0.266	0.236	0.203	0.225	0.226	0.262	/
STAAE	0.188	**0.234**	0.265	0.210	0.186	0.247	0.209	0.172	0.200	0.263	0.241	0.200	/
MAMSM	**0.221**	0.171	**0.321**	**0.320**	**0.274**	**0.262**	**0.302**	**0.288**	**0.213**	**0.307**	**0.256**	**0.293**	/
	L1	L2	S1	S2	R1	R2	M1	M2	M3	M4	M5	M6	Ave
SDL	0.209	0.177	0.272	0.273	0.244	0.201	0.133	0.146	0.122	0.143	0.146	0.206	0.202
STAAE	0.210	0.265	0.161	0.199	0.205	0.206	0.302	0.293	0.257	0.272	0.273	0.293	0.231
MAMSM	**0.305**	**0.272**	**0.352**	**0.374**	**0.374**	**0.258**	**0.343**	**0.345**	**0.299**	**0.297**	**0.314**	**0.322**	**0.295**

The bold values represent the maximum values of each column.

FIGURE 8
Comparison of FBNs obtained from SDL, STAAE, and MAMSM.

rate of the FBNs obtained from each method and GLM templates are shown in **Table 6**. We can see that the average OR value (0.295) of the brain network obtained by MAMSM is larger than that of STAAE (0.231) and SDL (0.202), which proves that the MAMSM proposed in this manuscript is superior to STAAE and SDL.

4. Discussion and conclusion

In this study, the multi-head attention mechanism and mask training method were applied to the analysis of tfMRI data, and a new loss function was constructed by task design curves for the mapping of functional brain networks. The multi-head attention mechanism helps the model better understand the situation where the same signal value in tfMRI signals may represent different states. Meanwhile, a mask training method was adopted to learn the relationship between the contexts of input sequences, and by combining a continuous mask and a discrete mask, deeper-level features were learned. The experimental results demonstrated that these techniques can improve the model's performance. By analyzing the comparison results of the intermediate features (attention-score, average-result) outputted from the model and the task design curves, it can be seen that the proposed model can better understand the tfMRI signals and the derived features are interpretable. The attention-score extracted after the model was trained represented the weight scores of different locations in each tfMRI sequence. The region with the highest score in the attention-score bears close resemblance to the area with the most significant alteration in the task design curves. The average-result obtained by simply sliding the attention-score achieved higher similarity with the task design curves than the results obtained by other methods.

We also leveraged prior knowledge (Task designs) to guide the model to learn the more efficient features, the task designs were introduced to build a new loss function which optimizes the model by cosine similarity error and MSE error. By analyzing the results, we found that this new loss function can improve the performance of the model. Other methods usually ignored the prior knowledge in their model, and experimental results show that MAMSM achieves better results than other methods when using the new loss function.

The experimental results show that the proposed method can achieve better generalization performance on smaller sample size, compared to other deep learning methods which require large amounts of data to achieve better results, such as STAAE (Dong et al., 2020b), ResAE (Dong et al., 2020a), Dvae (Qiang et al., 2020) and so on. Due to the characteristics of medical image data, such as high confidentiality and small sample size, the method proposed in this manuscript can have better development prospects in the future.

It is important to note that this study has certain limitations. Firstly, the relatively small size of the dataset employed may introduce noise when aggregating across groups, potentially impacting the outcomes of the brain network analyses. Furthermore, the present methodology places greater emphasis on temporal features of tfMRI data, and future investigations may benefit from incorporating a combination of convolutional neural network (CNN) models (Ronneberger et al., 2015; Liu et al., 2022) and visual transformer (VIT) models (Dosovitskiy et al., 2020; Liu et al., 2021) to extract spatial features, which may achieve better results. Additionally, the precise functional significance of some brain networks identified in the results is not fully understood at present, and hence, further research is warranted to explore the functional areas and meanings attributed to these networks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://humanconnectome.org/study/hcp-young-adult/document/q3-data-release.

Ethics statement

The studies involving human participants were reviewed and approved by the dataset is public and has been approved by its Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MH: methodology, formal analysis, software, visualization, writing—original draft, and writing—review and editing. XH: visualization and writing—original draft. EG, NQ, and XZ: writing—review and editing. ZW: software and visualization. ZK: validation and visualization. BG: conceptualization, methodology, writing—review and editing, funding acquisition, resources, and supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (no. 61976131) and the Fundamental Research Funds for Central Universities (JK202205022).

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

References

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage* 80, 169–189.

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

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage* 20, 1052–1063. doi: 10.1016/S1053-8119(03)00435-X

Calhoun, V. D., and Adali, T. (2012). Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* 5, 60–73. doi: 10.1109/RBME.2012.2211076

Canario, E., Chen, D., and Biswal, B. (2021). A review of resting-state fMRI and its use to examine psychiatric disorders. *Psychoradiology* 1, 42–53.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* arXiv: 1406.1078. doi: 10.3115/v1/D14-1179 [Preprint].

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., et al. (2021). "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proceedings of the 2021 IEEE automatic speech recognition and understanding workshop (ASRU)* (Cartagena: IEEE), 244–250. doi: 10.1109/ASRU51503.2021.9688253

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* arXiv: 1810.04805. [Preprint].

Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., and Li, Q. (2020b). "Spatiotemporal attention autoencoder (STAAE) for ADHD classification," in *Proceedings of the 23rd international conference, medical image computing and computer assisted intervention–MICCAI 2020* (Lima: Springer), 508–517. doi: 10.1007/978-3-030-59728-3_50

Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., and Li, Q. (2020a). "Discovering functional brain networks with 3D residual autoencoder (ResAE)," in *Proceedings of the 23rd international conference, medical image computing and computer assisted intervention–MICCAI 2020* (Lima: Springer), 498–507. doi: 10.1007/978-3-030-59728-3_49

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* arXiv: 2010.11929. [Preprint].

Ge, B., Makkie, M., Wang, J., Zhao, S., Jiang, X., Li, X., et al. (2016). Signal sampling for efficient sparse representation of resting state FMRI data. *Brain Imaging Behav.* 10, 1206–1222. doi: 10.1007/s11682-015-9487-0

Graves, A., Fernández, S., and Schmidhuber, J. (2005). "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proceedings of the 15th international conference, artificial neural networks: Formal models and their applications–ICANN* (Warsaw: Springer), 799–804.

Güçlü, U., and Van Gerven, M. A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Front. Comput. Neurosci.* 11:7. doi: 10.3389/fncom.2017.00007

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 16000–16009.

He, M., Hou, X., Wang, Z., Kang, Z., Zhang, X., Qiang, N., et al. (2022). "Multi-head attention-based masked sequence model for mapping functional brain networks," in *Proceedings of the 25th international conference, medical image computing and computer assisted intervention–MICCAI* (Singapore: Springer), 295–304. doi: 10.1007/978-3-031-16431-6_28

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., et al. (2017). Modeling task fMRI data via deep convolutional autoencoder. *IEEE Trans. Med. Imaging* 37, 1551–1561. doi: 10.1109/TMI.2017.2715285

Jiang, X., Yan, J., Zhao, Y., Jiang, M., Chen, Y., Zhou, J., et al. (2023). Characterizing functional brain networks via spatio-temporal attention 4D convolutional neural networks (STA-4DCNNs). *Neural Netw.* 158, 99–110. doi: 10.1016/j.neunet.2022.11.004

Jiang, X., Zhang, T., Zhang, S., Kendrick, K. M., and Liu, T. (2021). Fundamental functional differences between gyri and sulci: Implications for brain function, cognition, and behavior. *Psychoradiology* 1, 23–41. doi: 10.1093/psyrad/kkab002

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., and Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 26, 317–329. doi: 10.1016/j.neuroimage.2005.01.048

Lee, Y.-B., Lee, J., Tak, S., Lee, K., Na, D. L., Seo, S. W., et al. (2016). Sparse SPM: Group sparse-dictionary learning in SPM framework for resting-state functional connectivity MRI analysis. *Neuroimage* 125, 1032–1045. doi: 10.1016/j.neuroimage.2015.10.081

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision* (Montreal, QC: IEEE), 10012–10022. doi: 10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 11976–11986. doi: 10.1109/CVPR52688.2022.01167

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., et al. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Med. Image Anal.* 20, 112–134.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., et al. (2014). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Trans. Biomed. Eng.* 62, 1120–1131. doi: 10.1109/TBME.2014.2369495

McKeown, M. J. (2000). Detection of consistently task-related activations in fMRI data with hybrid independent component analysis. *Neuroimage* 11, 24–35. doi: 10.1006/nimg.1999.0518

Mourao-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., and Brammer, M. (2006). The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. *Neuroimage* 33, 1055–1065. doi: 10.1016/j.neuroimage.2006.08.016

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Power, J. D., Fair, D. A., Schlaggar, B. L., and Petersen, S. E. (2010). The development of human functional brain networks. *Neuron* 67, 735–748. doi: 10.1016/j.neuron.2010.08.017

Qiang, N., Dong, Q., Ge, F., Liang, H., Ge, B., Zhang, S., et al. (2020). Deep variational autoencoder for mapping functional brain networks. *IEEE Trans. Cogn. Dev. Syst.* 13, 841–852. doi: 10.1109/TCDS.2020.3025137

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th international conference, medical image computing and computer-assisted intervention–MICCAI* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shen, H., Xu, H., Wang, L., Lei, Y., Yang, L., Zhang, P., et al. (2017). Making group inferences using sparse representation of resting-state functional mRI data with application to sleep deprivation. *Hum. Brain Mapp.* 38, 4671–4689. doi: 10.1002/hbm.23693

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., and Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv* arXiv: 2104.06644. doi: 10.18653/v1/2021.emnlp-main.230 [Preprint].

Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F. (2014). Group-PCA for very large fMRI datasets. *Neuroimage* 101, 738–749. doi: 10.1016/j.neuroimage.2014.07.051

Sporns, O., and Betzel, R. F. (2016). Modular brain networks. *Annu. Rev. Psychol.* 67, 613–640. doi: 10.1146/annurev-psych-122414-033634

Thirion, B., and Faugeras, O. (2003). Dynamical components analysis of fMRI data through kernel PCA. *Neuroimage* 20, 34–49. doi: 10.1016/S1053-8119(03)00316-1

Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv* arXiv: 2203.12602. [Preprint].

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: An overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., et al. (2022). "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (New Orleans, LA: IEEE), 9653–9663.

Yan, J., Chen, Y., Xiao, Z., Zhang, S., Jiang, M., Wang, T., et al. (2022). Modeling spatio-temporal patterns of holistic functional brain networks via multi-head guided attention graph neural networks (Multi-Head GAGNNs). *Med. Image Anal.* 80:02518. doi: 10.1016/j.media.2022.102518

Zhang, S., Li, X., Lv, J., Jiang, X., Guo, L., and Liu, T. (2016). Characterizing and differentiating task-based and resting state fMRI signals via two-stage

sparse representations. *Brain Imaging Behav.* 10, 21–32. doi: 10.1007/s11682-015-9359-7

Zhang, W., Lv, J., Li, X., Zhu, D., Jiang, X., Zhang, S., et al. (2018). Experimental comparisons of sparse dictionary learning and independent component analysis for brain network inference from fMRI data. *IEEE Trans. Biomed. Eng.* 66, 289–299. doi: 10.1109/TBME.2018.2831186

Zhao, Y., Li, X., Zhang, W., Zhao, S., Makkie, M., Zhang, M., et al. (2018). "Modeling 4d fMRI data via spatio-temporal convolutional neural networks (ST-CNN)," in *Proceedings of the 21st international conference, medical image computing and computer assisted intervention–MICCAI 2018* (Granada: Springer), 181–189.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., et al. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv* arXiv: 2111.07832. [Preprint].

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 17 May 2023
DOI 10.3389/fnins.2023.1152619

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Sara Ponticorvo,
University of Salerno, Italy
Yuanqiang Zhu,
Fourth Military Medical University, China

*CORRESPONDENCE
Minghao Dong
✉ dminghao@xidian.edu.cn
Chenwang Jin
✉ jin1115@xjtu.edu.cn

†These authors have contributed equally to this work

RECEIVED 28 January 2023
ACCEPTED 26 April 2023
PUBLISHED 17 May 2023

CITATION
Wang H, Yao R, Zhang X, Chen C, Wu J, Dong M and Jin C (2023) Visual expertise modulates resting-state brain network dynamics in radiologists: a degree centrality analysis. *Front. Neurosci.* 17:1152619. doi: 10.3389/fnins.2023.1152619

Visual expertise modulates resting-state brain network dynamics in radiologists: a degree centrality analysis

Hongmei Wang[1,2†], Renhuan Yao[3†], Xiaoyan Zhang[4], Chao Chen[5], Jia Wu[6], Minghao Dong[4,7*] and Chenwang Jin[1*]

[1]Department of Radiology, First Affiliated Hospital of Xi'an, Jiaotong University, Xi'an, China, [2]Department of Medical Imaging, Inner Mongolia People's Hospital, Hohhot, China, [3]Department of Nuclear Medicine, Inner Mongolia People's Hospital, Hohhot, China, [4]Engineering Research Center of Molecular and Neuro Imaging of Ministry of Education, School of Life Science and Technology, Xidian University, Xi'an, China, [5]PLA Funding Payment Center, Beijing, China, [6]School of Foreign Languages, Northwestern Polytechnical University, Xi'an, Shaanxi, China, [7]Xi'an Key Laboratory of Intelligent Sensing and Regulation of Trans-Scale Life Information, School of Life Science and Technology, Xidian University, Xi'an, China

Visual expertise reflects accumulated experience in reviewing domain-specific images and has been shown to modulate brain function in task-specific functional magnetic resonance imaging studies. However, little is known about how visual experience modulates resting-state brain network dynamics. To explore this, we recruited 22 radiology interns and 22 matched healthy controls and used resting-state functional magnetic resonance imaging (rs-fMRI) and the degree centrality (DC) method to investigate changes in brain network dynamics. Our results revealed significant differences in DC between the RI and control group in brain regions associated with visual processing, decision making, memory, attention control, and working memory. Using a recursive feature elimination-support vector machine algorithm, we achieved a classification accuracy of 88.64%. Our findings suggest that visual experience modulates resting-state brain network dynamics in radiologists and provide new insights into the neural mechanisms of visual expertise.

KEYWORDS

degree centrality, visual expertise, object recognition, support vector machine, radiologist

Introduction

Visual expertise is a cognitive process that involves visual object recognition ability in a specific domain, resulting in superior visual object recognition performance (Harel, 2016; Wang et al., 2021). The development of visual expertise is thought to involve reciprocal interactions between the visual system and multiple high-level areas across the brain (Harel, 2016). In particular, visual expertise is essential for the development of radiological expertise, which enables radiologists to rapidly and accurately recognize abnormalities in medical images (Haller and Radue, 2005; Harley et al., 2009; Hendee, 2010; Melo et al., 2011). However, the neural mechanisms underlying visual expertise in radiology remain poorly understood, particularly with regard to resting-state brain network dynamics. Resting-state functional magnetic resonance imaging (rs-fMRI) can be used to investigate the intrinsic activity of multiple neural networks simultaneously and may help uncover the neural basis of visual expertise in radiology. In this study, we used rs-fMRI to investigate how visual expertise induced by experience modulates the dynamics of brain networks in radiologists.

Previous neuroimaging studies have investigated the neural mechanisms underlying visual expertise using different expertise models. Martens et al. (2018) found that bird expertise-related neural changes involved both low-level and high-level visual regions as well as frontal lobe areas, suggesting that expertise can modulate neural correlates that are specific to the domain as well as those that are more general. Similarly, research on London taxi drivers by Spiers and Maguire (2006) revealed widespread patterns of activation along visual pathways and other brain regions such as the parahippocampal cortex, retrosplenial cortex, and prefrontal structures, indicating their association with scene processing, navigation, and spatial processing when participants inspected landmark objects in city scenes. In the context of radiological expertise, previous studies have reported selective activations in the brain regions of radiologists such as the bilateral middle frontal gyrus (MFG) and left superior frontal gyrus (SFG), which are linked to visual attention and memory retrieval, when comparing brain responses to radiological images between radiologists and laypersons (Haller and Radue, 2005). Furthermore, it was found that the fusiform face area (FFA) was more active when radiologists viewed domain-related images and contributed to the recognition of normal anatomical features based on subjective similarity rather than physical similarity (Harley et al., 2009). This finding was supported by Bilalic et al. (2016) who showed that FFA could help radiologists discriminate X-ray stimuli from other stimuli and then contribute to the evaluation of radiographic images. Lastly, Wang et al. (2021) proposed that visual experience could modulate the functional adaptation of the visual cortex and other cognitive areas that are responsible for decision making, semantic knowledge, and attention, as evidenced by widely altered functional connectivity in the entire cortex including the SFG, MFG, orbitofrontal cortex (OFC), and fusiform gyrus (FuG). Collectively, these studies suggest that the activation of these circuits or brain areas constitutes a cortical organizing principle of visual expertise in the brain, such as visual processing, attention control, decision making, and semantic memory.

Radiology is a particularly suitable domain for investigating the impact of visual experience on expertise because it allows for a comparison between experienced radiologists or medical interns and lay persons who lack experience, enabling the identification of distinguishing traits (Bilalic et al., 2016). Functional magnetic resonance imaging (fMRI) is a promising method to uncover functional adaptations in the entire brain cortex associated with visual expertise. Resting-state brain activity refers to the intrinsic response of the brain in the absence of thinking activity (Smitha et al., 2017; Canario et al., 2021) and the observed brain activity is regarded as being responsible for coding prior experience (Albert et al., 2009; Dong et al., 2014). However, few studies have utilized resting-state fMRI (rs-fMRI) to investigate the neural mechanisms of visual expertise in radiologists. Degree centrality (DC) is a graph-based measurement that can reveal the network dynamics modified by prior experience and node centrality for visual expertise (Reynolds et al., 2018; Liu and Lai, 2022). A support vector machine (SVM) is a machine learning-based pattern classification approach that has unique advantages in understanding small sample learning problems and has been widely applied in biological data processing (Cherkassky, 1997; Li et al., 2014; Liu et al., 2014). The most discriminatory parts of the brain based on SVM represent the most striking feature between the two groups and reveal underlying expertise-related neurobiology (Ding et al., 2015; Gao et al., 2022). By utilizing rs-fMRI, DC, and SVM, we aim to gain a deeper understanding of the neural mechanisms of visual expertise in radiologists.

The main goal of this study was to explore how visual experience modulates DC in resting-state activity and to understand the neural correlates of visual expertise using a model of radiologists ($n = 22$) and rs-fMRI. The DC method combined with a novel but sensitive machine learning method, i.e., a recursive feature elimination-support vector machine (RFE-SVM) (Ding et al., 2015), was employed to look for the highest discriminative power between the radiology intern (RI) group and the normal control (NC) group. We expect that visual experience modulates the expertise-related brain areas beyond the visual cortex and even other cognitive areas, thus supporting working memory (WM), memory, attention control, and decision making (Harel et al., 2013; Harel, 2016; Wang et al., 2021).

Materials and methods

All study procedures were approved by the Subcommittee on Human Studies of the First Affiliated Hospital of Medical College in Xi'an Jiaotong University and were conducted in accordance with the Declaration of Helsinki.

Participants

Twenty-two radiology interns and 22 matched subjects were recruited in our study. All of the subjects in the RI group were undergraduates majoring in radiology who interned at the First Affiliated Hospital of Xi'an Jiaotong University. Before rotation, all of the participants received basic medical education at their college. The RI group had X-ray department rotation experience, mainly in interpreting X-ray images for 4 weeks, during which time they practiced 6 days per week and read 25–35 cases per day. The total length of training was 26 ± 2.4 (mean ± standard deviation, SD) days. Scrutinizing the images displayed on the screen and completing the X-ray reports were the main tasks of every intern's training. Each of the interns had a senior tutor providing basic clinical support. After 4 weeks of rotation, at least 600 reports written by each RI were recorded in the Picture Archiving and Communication System (PACS), which were modified by the instructor to meet the "degree of agreement" requirements. The subjects in the NC group were from other majors and had never participated in any form of medical imaging training nor received any related education. The average ages of the RI group and NC group were 23 ± 0.7 years and 23 ± 0.5 years, respectively. The sex distribution in the two groups was the same (11 males; 11 females). The recruitment criteria of all subjects included the following: (1) the participants were physically healthy and right-handed; (2) the subjects and their immediate family members had no past or present neurological, psychiatric, or neuropsychological disorders and had no history of head trauma or brain tumor by

medical history, physical, and neurological examinations; and (3) participants took no relevant drugs before or during the internship. Written consent forms were obtained from all the participants.

Behavioral measurement

Both the RI group and NC group completed the same behavioral tasks. We conducted the prescreening tasks using a face-to-face questionnaire to exclude confounding factors, such as visual expertise from other domains (e.g., cars, chess, birds, and mushrooms). The subjects' behavioral test of the visual expertise level was restricted to X-ray films because of the high specialty for required perceptual expertise (Nakashima et al., 2015). Participants in the RI group were required to pass a practical examination about radiological anatomy and interpretation of X-ray films to verify that they had reached a required level of expertise. The Cambridge Face Memory Test (CFMT) and Radiological Expertise Task (RET) were employed to measure face expertise and radiological expertise in our study. The RET consists of 100 standard X-ray images of adults including 65 positive images and 35 negative images, from the PACS of the X-ray image bank under the guidance of three senior independent expert radiologists with more than 10 years of radiological experience and who were proficient in reading X-ray images. The three senior experts not only scrutinized the pathological appearance of the selected films and confirmed the approval of the reports but also evaluated the level of difficulty of the reports on a scale of 1–3. Sixty-five positive X-ray images contained one nodule without any other conclusions in the corresponding reports and 35 negative images were normal X-rays without any lesions. The level of difficulty for grades 1–3 in all 100 images used in the RET accounted for 55%, 30%, and 15% of the images, respectively. The detailed procedures of CFMT and RET were introduced in our previous research (Zhang et al., 2022).

MRI data acquisition

fMRI data were collected from 8:30 a.m. to 12:30 a.m. to eliminate the time-of-day effect (Hasler et al., 2014). Brain imaging scans were performed on a 3T GE scanner (EXCITE; General Electric; Milwaukee; Wisc.) at the imaging center of Xi'an Jiaotong University First Affiliated Hospital. A standard birdcage head coil and restraining foam pads were used to minimize head motion and protect participants' hearing. Resting-state functional images were acquired by an echo-planar-imaging sequence, and the specific parameters included 32 contiguous slices with a slice thickness = 4 mm, layer interval = 0, TR = 2,000 ms, TE = 30 ms, FA = 90°, FOV = 240 mm × 240 mm, data matrix = 64 × 64, voxel size = 3.75 mm × 3.75 mm × 4 mm, total volumes = 190, and scanning time = 380s. During the entire resting process, the subjects had to keep their eyes closed, stay awake, and try to keep their minds blank without having any particular thoughts. Additionally, an MPRAGE T1-magnetization high resolution anatomical image (1 × 1 × 1 mm) was acquired for each participant with the following parameters: TE = 2.26 ms, TR = 1,900 ms, flip angle = 9°, FOV = 256 mm, slice thickness = 1 mm, and matrix = 256 × 256. A total of 176 slices in the sagittal orientation were acquired. Potential clinical abnormalities of each participant were assessed by two expert radiologists based on the structural images. No participants were excluded at this level.

Resting-state fMRI preprocessing

Statistical Parametric Mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm) and the Data Processing Assistant for Resting-State fMRI (DPARSF 4.5, http://rfmri.org/DPARSF) were used for MRI data preprocessing. The preprocessing steps were as follows: (1) DICOM data were converted to NIFTI format; (2) the first 10 time points were removed for stability of the magnetic field and to allow the subjects to adapt to the experimental environment; (3) slice time correction was conducted for the remaining time points; (4) motion correction was carried out using rigid body transformation to fix the brain at the same target position; (5) the functional images were coregistered to the subject's anatomical images, and all the processed data were divided into gray matter, white matter, and cerebrospinal fluid by the exponentiated lie algebra (DARTEL) tool (Ashburner, 2007); (6) a higher-level Friston-24 model was employed to regress out head motion; (7) the nuisances such as global signal, white matter signal, and cerebrospinal fluid signal were regressed; (8) all the resting functional images were normalized to MNI space using the deformation field maps obtained from structural image segmentation; (9) the normalized fMRI data were resampled to 3 mm isotropic voxels; (10) the images were then spatially smoothed with a 6 mm full width at half maximum (FWHM) Gaussian kernel; and (11) linear trend removal and temporal bandpass filtering (0.01–0.08 Hz) were performed to reduce the effect of low-frequency drifts and high-frequency noise.

Feature extraction

Generation of voxel wise and region wise DC maps

The DC index has unique superiority (i.e., high sensitivity, specificity, and reliability) in reflecting the dynamics of brain networks (Zuo and Xing, 2014). In current study, the DC method was employed to look for the neuroimaging features between groups. The specific steps were as follows: first, the BOLD time course of each voxel was extracted and its Pearson's correlation with all other voxels in the whole brain was analyzed. Every voxel with positive correlation coefficients >0.2 was selected, which can eliminate the weak correlation due to signal noise to ensure that voxels have higher regional functional connectivity strength values. Fisher's r-to-z transformation was conducted to derive the Z score matrix and improve normality for the resulting voxel for each participant. Then, the DC value of each subject was divided by the mean of the whole brain to achieve standardization, which can eliminate individual differences. The DC map of the whole brain based on the voxel-level data was obtained. After that, the voxel wise DC map was averaged into a region wise DC map. The Brainnetome atlas was employed to divide the DC map into 246 regions of interest (ROIs) (Fan et al., 2016). The DC values of all the ROIs were averaged to obtain the average DC value of each region.

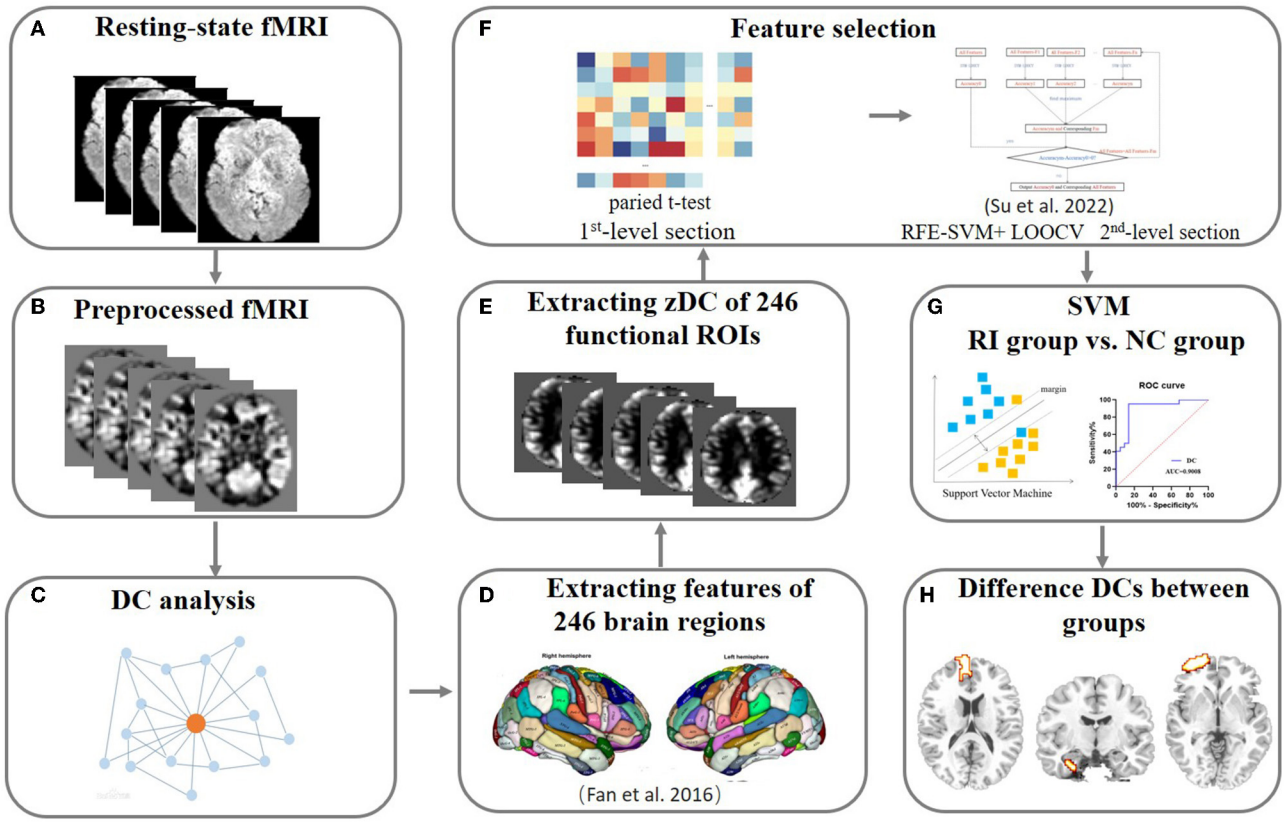


FIGURE 1
The pipeline of the rs-MRI data analysis. **(A–C)** The resting-state MRI data were collected and preprocessed following procedures described in the Methods. Then, the DC for each voxel was calculated and used for future feature selection. **(D–F)** Feature selection. Two-step feature selection was performed and the first level used a two-sample approach to perform the regional average feature. Then, RFE-SVM modeling with LOOCV was employed to search for the most remarkable features between groups. **(G, H)** SVM modeling. Reliable SVM classification results and the brain areas with robust differences in DC values between groups were obtained to reflect the alteration of dynamics in the whole-brain network. rs-MRI, resting-state MRI; fMRI, functional magnetic resonance imaging; DC, degree centrality; RFE-SVM, recursive feature elimination-support vector machine; SVM, support vector machine; LOOCV, leave-one-out cross-validation.

Finally, the mean DC values from the 246 ROIs then served as the input vector for the classification procedure.

Feature selection

Feature selection is a hotspot in bioinformatics and is critical in medical studies. Its process is to extract informative features from complex high-dimensional data (Du et al., 2017). We performed a two-stage feature selection procedure in our study. Firstly, we used a two-sample t test to identify the differences in the region wise DC maps between the two groups in a leave-one-out fashion, with a threshold of $p < 0.05$ considered significant. The resulting region wise features were then used in the second-level elimination. In our study, the recursive feature elimination-support vector machine (RFE-SVM) Guyon et al. (2002) is employed for the purpose of feature selection that combines recursive feature elimination with SVM modeling. Basically, we used the RFE-SVM approach in a leave-one-out cross-validation (LOOCV) framework to recursively eliminate the least useful features until further elimination resulted in reduced accuracy. The basic idea behind RFE-SVM is introduced as follows: in each iteration, the contribution to classification accuracy is determined by eliminating one feature at a time using SVM-LOOCV. Then, the features with zero contribution to classification accuracy is taken away from feature set which is to be used as input for next round of iteration. These steps are repeated until the number of features reaches zero. The feature set with highest classification accuracy is used as the outcome of RFE-SVM and sent to SVM for modeling. For this step, we used several performance indicators, including accuracy, sensitivity, specificity, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC), to evaluate the efficiency of the RFE-SVM classifier. LOOCV was also used to validate the model. Note that a linear SVM classifier model with a soft interval separation and hinge loss function, as it is commonly used in neuroimaging data and produces interpretable results (Rasmussen et al., 2011). The pipeline of rs-MRI data and feature selection processing in this study is illustrated in Figure 1.

Correlation analysis

To evaluate the relationship between behavioral measurements and the dynamics of the resting brain network in the two groups,

voxel wise Pearson's correlation analysis was conducted between the averaged DC values and outcome of behavioral tasks (i.e., RET scores and response times). The significance level was set at $p < 0.05$ after multiple comparison correction (false discovery rate, FDR).

Results

There were no significant differences in age or sex between the groups ($p > 0.05$). The mean practice level duration and cases in the total RI group are shown in Table 1.

Results of behavioral tests

The behavioral performance of the RI and NC groups is summarized in Figure 2 and Table 1. Compared with the NC group, the RI group had significantly higher RET scores, indicating that the visual experience enabled the RI group to have better nodule recognition ability than the NC group ($p < 0.001$, Mann–Whitney U-test). The response time of RET in the RI group was much shorter than that in the NC group, suggesting that the RI group can recognize nodules much faster than the NC group ($p < 0.001$, Mann–Whitney U-test). There was no significant difference in CMFT scores between the two groups, which demonstrated that the two groups had similar face recognition abilities ($p > 0.05$, Mann–Whitney U-Test).

TABLE 1 Demographic data of the radiological intern group and normal control group.

Labels	Radiologists ($n = 22$) Mean ± SD	Controls ($n = 22$) Mean ± SD	p-values
Length of training	26 ± 2.4	N/A	–
Cases in total	767.4 ± 82.6	N/A	–
RET*^	0.80 ± 0.04	0.53 ± 0.04	<0.001
Response time of RET (s)*	2.6 ± 0.4	3.7 ± 0.7 s	<0.001
Face expertise	56.95 ± 5.23	58.68 ± 5.31	0.28

*Denotes significant difference between groups ($p < 0.001$).
^Denotes that the Mann–Whitney test was used.
SD, standard deviation; s, seconds; RET, radiological expertise task; CFMT, Cambridge face memory test.

SVM classification results

The iteration procedure of feature selection based on RFE-SVM is presented in Figure 3A. The highest classification accuracy was observed in the seventh subset. The brain regions with discriminative power included the bilateral SFG, left MFG, right orbital gyrus (OrG), left FuG, and bilateral parahippocampal gyrus (PhG). The details of the brain regions are shown in Table 2 and Figure 4. The SVM classification accuracy was 88.64%, sensitivity was 81.82%, specificity was 95.45%, and AUC was 0.9008. The ROC curve of classification accuracy for RFE-SVM is presented in Figure 3B.

Results of correlation analysis

A significant positive correlation between the average voxel wise DC of the left FuG and the level of radiological expertise (i.e., RET scores) was found in the RI group after multiple comparisons ($r = 0.51$, $p < 0.05$, Figure 5). No significant correlations were found between other indicators of the behavioral tests and DC in the RI or the NC groups.

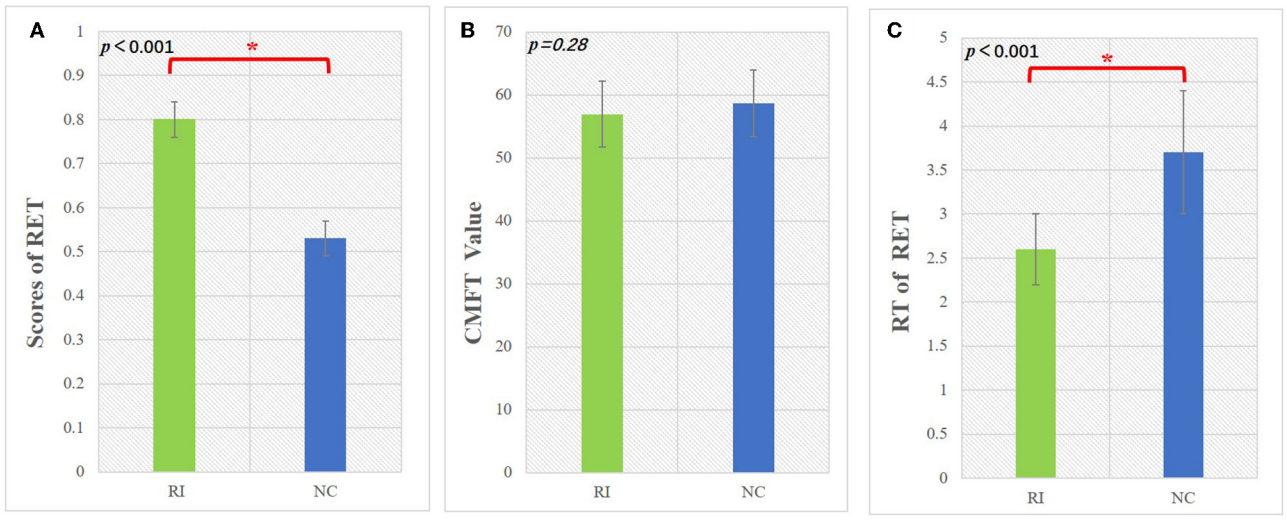


FIGURE 2
The performance of the behavioral test in the RI group and NC group. **(A)** The scores of lung nodule identification for the two groups. **(B)** The different values in the CFMT for each group. **(C)** The response time of recognizing the lung nodule for each group. RET, radiological expertise task; CFMT, Cambridge face memory test; RT, response time; RI, radiology intern; NC, normal control. *Indicates significant group differences ($p<0.05$).

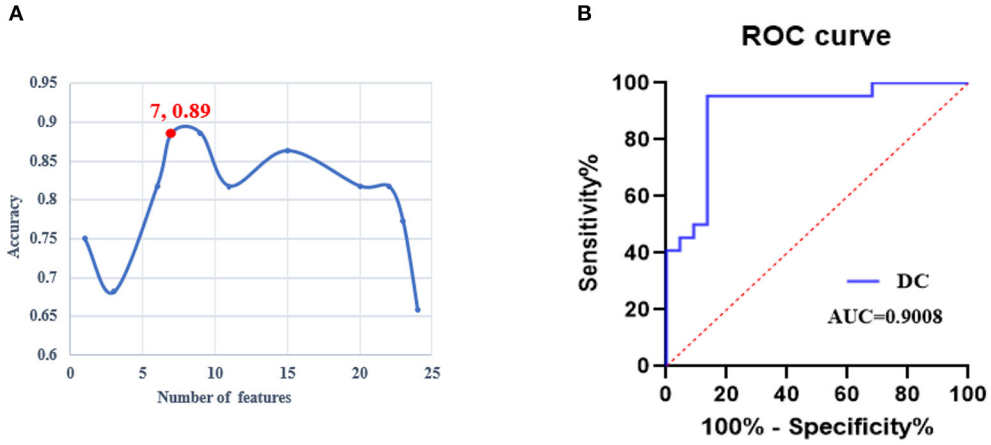


FIGURE 3
The performance of the RFE-SVM classifier. **(A)** The iteration procedure of feature selection based on RFE-SVM. **(B)** The receiver operating characteristic curve of the RFE-SVM classifier. ROC, receiver operator characteristic; DC, degree centrality; AUC, area under the curve; RFE-SVM, recursive feature elimination support vector machine.

TABLE 2 The difference in DC values between the RI and NC groups.

Cognitive component	Brain region	Subregions	Brodmann's areas	Side	Weight
Attention control	MFG	MFG_L_7_7	BA10 (lateral)	L	−0.59
Decision making	OrG	OrG_R_6_4	BA11(medial)	R	−0.60
Visual processing	FuG	FuG_L_3_3	BA37 (ventral and lateral)	L	0.56
Memory	PhG	PhG_L_6_1	BA35/36	L	−0.63
		PhG_R_6_4	A28/34	R	−0.99
Working memory	SFG	SFG_R_7_6	BA9 (medial)	R	−0.66
		SFG_L_7_7	BA10 (medial)	L	0.64

BA, brodmann area; FuG, fusiform gyrus; MFG, middle frontal gyrus; OrG, orbital gyrus; PhG, parahippocampal gyrus; SFG, superior frontal gyrus; L, left; R, right.

Discussion

Visual expertise is a complex skill that requires learning from a vast amount of domain-specific visual information (Dong et al., 2022). Several studies have examined the neural mechanisms underlying radiologists' expertise, identifying high-order cognitive and low-order visual factors such as visual processing, WM, attention control, and decision making as crucial components (Donovan and Litchfield, 2013; Harel, 2016; Annis and Palmeri, 2019). However, the extent to which visual experience modulates resting-state brain activity in radiologists remains unclear. This study aimed to address this gap by investigating how real-world visual experience affects the DC values of resting-state brain activity in radiologists. Our behavioral results showed that the RI group performed better after training than the NC group (Figure 2), and the imaging data analysis demonstrated that seven brain subregions in the visual cortex, prefrontal lobe, and limbic system had the highest discriminative power in between-group comparisons (Figure 4 and Table 2). These results were obtained using RFE-SVM, which demonstrated excellent classification efficiency with high accuracy, sensitivity, and specificity (Figures 3A, B). Additionally, we found a significant positive correlation between RET scores and the DC values of the left FuG, indicating that the functional connectivity of this region is related to visual expertise (Figure 5). To our knowledge, this study is the first to investigate DC level changes in radiologists' resting brains in response to real-world visual experience. The results provide new insights into the neural mechanisms underlying visual expertise, and the findings may have practical applications for radiologist training. Overall, our study highlights the importance of considering resting-state brain activity in understanding how visual expertise develops and may help inform future research in this area.

The increased DC level of the left FuG in radiologists

Compared with the NC group, the RI group had increased DC values in the left FuG which controls visual processing (Figure 4 and Table 2). Additionally, we found a significant positive correlation between the DC value of the left FuG and RET scores in the RI group (Figure 5). The acquisition of visual experience may be accompanied by functional enhancement of visual processing supporting radiologists' superior performance (Haller and Radue, 2005; Wang et al., 2021). The FuG, located in the human ventral temporal cortex (VTC), is a pivotal functional brain module

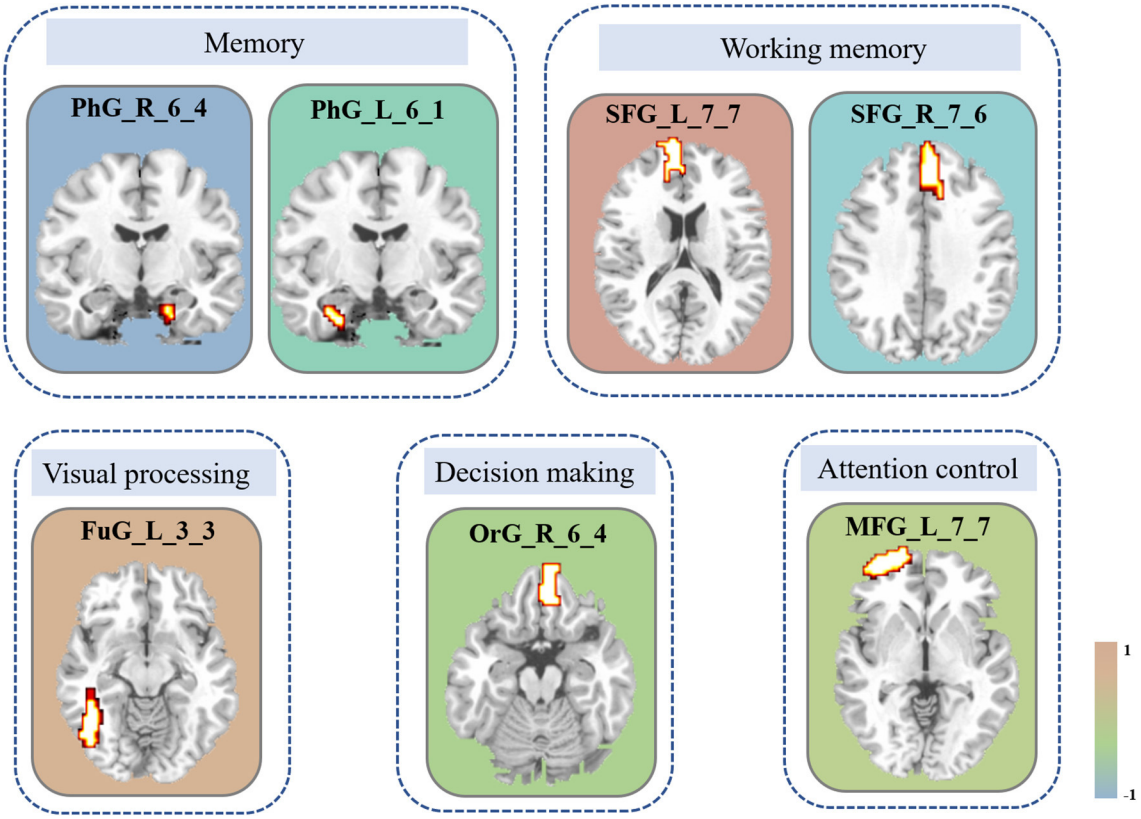


FIGURE 4
Brain areas with the most discriminative ability between groups. The constitutional diagram is categorized by visual and cognitive components. The color bar shows the size of the weight. Note that the positive direction represents the increased DC values and vice versa. MFG, middle frontal gyrus; OrG, orbital gyrus; FuG, fusiform gyrus; PhG, parahippocampal gyrus; SFG, superior frontal gyrus; DC, degree centrality.

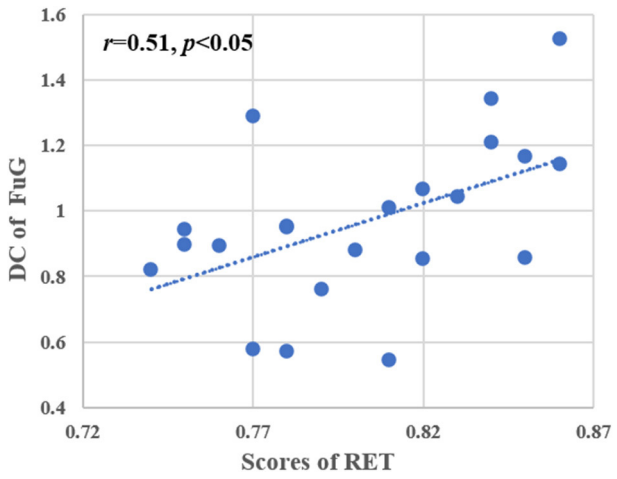


FIGURE 5
Correlation analysis between RET scores and DC values of the left FuG. Pearson correlation was used to assess significance ($p < 0.05$, multiple comparison corrected). RET, radiological expertise task; FuG, fusiform gyrus; DC, degree centrality.

within the high-level visual cortex (Weiner and Zilles, 2016) which is a key-structure in high-level visual processing for object recognition (Grill-Spector et al., 2001). The FuG contains several category-selective regions for the recognition of different visual stimuli, including the FFA (Kanwisher et al., 1997), fusiform body area (Peelen and Downing, 2005), and visual word-form area (Cohen et al., 2000). Several neuroimaging studies using visual expertise models, such as cars (McGugin et al., 2015), chess (Bilalić, 2016), faces (Goold and Meng, 2017), and radiology (Haller and Radue, 2005) reported activation of the FuG in task fMRI studies. Specifically, the right FuG was engaged in the processing of non-face expertise visual stimuli (Xu, 2005; Harley et al., 2009; Engel et al., 2009) and mediated the formation of category-specific representations (van der Linden et al., 2008). Moreover, the right FFA plays an important role in visual discrimination that can be fine-tuned by experience with other domain categories (Engel et al., 2009). Furthermore, previous studies have consistently reported that the left FuG plays a more prominent role than the right FuG in processing non-face related information (Devlin et al., 2006; Bi et al., 2014; Bilalic et al., 2016). In details, the left FuG was engaged in visual word recognition as a connector between the abstract visual information and higher order properties of the stimulus (Devlin et al., 2006) and not only participated in visual categorization learning but also its activity could be modulated by visual learning (Goold and Meng, 2017). In radiological expertise, left FuG activation made the radiologists more sensitive to radiological images and reliably distinguished between upright

and inverted X-rays (Bilalic et al., 2016). Additionally, a prior study found that the activity of the left FuG was positively correlated with participants' perceptual performance (Bi et al., 2014) supporting the pivotal role of the FuG in supporting recognition efficiency (Zhang et al., 2022).

In the current study, we speculated that the left FuG plays a vital role in recognizing the stimuli of radiological images, Furthermore, the short-term extraordinarily high load and repetitive usage of the visual system by the RI group can modify the visual processing of radiological stimuli. The fine-tuned behavioral performance and functional adaptation manifest in the superior ability of recognizing the nodule stimulus and stronger neural reflections in the resting state to make the brain much more efficient in detecting nodule-specific features.

The decreased DC level of the right OrG in radiologists

Decreased DC values of the right OrG were found in the RI group compared with those of the NC group (Figure 4 and Table 2). The OrG, as an OFC subregion (Rudebeck and Rich, 2018), is responsible for decision-making by primarily adjusting the utilities associated with different sensory stimuli (Lee et al., 2007) and plays a critical role in flexible, outcome-guided behavior (Liu et al., 2020). Decision-making is the process of choosing a particular response and further flexibly modifying cognitive and sensorimotor operations based on an evaluation of potential costs (Lee et al., 2007), which is necessary for expert visual processing. Decision making is part of the object recognition process during image interpretation (Wang et al., 2021). Of note, decision-making ability changes dynamically and continually as experience increases (Lee et al., 2007). Hence, the OFC was activated when the participants faced low-cost situations, such as either passively viewing information or selecting among options (Volz et al., 2006). A previous study on baseball batter expertise also verified that the OFC was responsible for expertise-driven rapid visual decisions (Muraskin et al., 2015). In Kirk's study on aesthetic expertise, the recruitment of the OFC between experts and non-experts suggested that this region was involved in expertise-related reward processing (Kirk et al., 2009). A study based on a chess model reported that the OFC appeared to be activated in this comparison between experts and novices (Krawczyk et al., 2011). In the current study, we propose that visual experience modulates radiologists' decision-making processes. Specifically, when radiologists face domain-specific options, they need to make decisions by employing many brain resources to recognize radiological stimuli.

The changed DC level of bilateral SFG in radiologists

Compared with those of the NC group, changed DC values of the bilateral SFG were found in the RI group (Figure 4 and Table 2). Multiple previous studies have shown that the SFG plays important roles in WM (Klingberg et al., 1997; Su et al., 2022). WM is a central mental capacity; it provides the platform for holding and manipulating thoughts and for organizing goal-directed behavior (Miller et al., 2018). WM capacity, which refers to the ability to retain the maximum amount of information, is a vital factor for problem solving and reasoning ability (Westerberg and Klingberg, 2007). The acquisition of visual expertise might improve WM performance (Moore et al., 2006). The neuroimaging study of Haller and Radue (2005) found that the enhanced neuronal activations of the SFG manifested in better WM capability in the process of radiological expertise. Kesler et al. (2011) found significantly increased activation of the SFG in visual tasks, which participated in online monitoring and manipulation of task-related information. Ouellette et al found lower activation of the lateral SFG in trained radiologists while they viewed medical images, suggesting that WM is a crucial component of radiology expertise and more efficient in radiologists (Ouellette et al., 2020). In our current study, different trends in the bilateral SFG showed that the increased DC values in the left SFG and decreased DC values in the right SFG were closely associated with WM when utilizing radiological expertise. Taking the weight of the brain area into consideration, the overall trend of DC values tended to be negative in the right SFG. Therefore, the decreased DC values of the right SFG may demonstrate increased neural efficiency of the WM process, thus enabling the RI group to spend less energy making a judgment and obtaining a good result compared with that of the NC group. Furthermore, we propose that the altered dynamics of the brain network when acquiring radiological expertise might support remodeling of the WM process reflecting more automated encoding and maintenance WM capacity, indicating a more efficient mechanism subserving visual expertise.

The decreased DC level of left MFG in radiologists

Decreased DC values of the left MFG in the RI group were found compared with values of the NC group in our study (Figure 4 and Table 2). A previous neuroimaging study found that the MFG participated in visual attention based on the model of radiologists (Haller and Radue, 2005). Selective attention can optimize the processing of information, make radiologists rapidly search for a particular "target" in radiographic images and adjust their response to information collected and compared to previously learned reference images (Haller and Radue, 2005; Harley et al., 2009). Attention has an important impact on visual expertise, even in the earliest step of visual processing (Harel, 2016). The left MFG has a crucial role in the dorsal attention network (DAN) and ventral attention network (VAN) to facilitate interactions between the two networks during attentional processing (Briggs et al., 2021). It has been found that the MFG is an important center facilitating attentional processes (Japee et al., 2015). Haller and Radue (2005) found enhanced neuronal activation of the MFG in radiologists compared with that in non-radiologists, suggesting that the MFG participated in the process of radiological expertise and played an important role in attention control. In contrast, the study of Melo et al. (2011) reported lower activation of the MFG in radiologists than non-radiologists when they were observing medical images. The evidence summarized above suggested that

short-term experience could adjust the process of attention control to make it more efficient and enable trainees to have more flexibility in manipulating limited attentional resources so that residual resources could be allocated validly to other brain regions supporting more demanding tasks.

The decreased DC level of bilateral PhG in radiologists

The comparison between groups revealed decreased DC values of the bilateral PhG in the RI group compared with those of the NC group (Figure 4 and Table 2). The PhG is an important center for memory processing (Lin et al., 2021). The acquisition of visual expertise might be accompanied by the alteration of memory representations (Annis and Palmeri, 2019). Existing neuroimaging studies have reported that chess experts and expert archers recruited more activation of the PhG when responding to domain-specific stimuli (Bilalić et al., 2010; Kim et al., 2011). In our study, the deceased DC values of the bilateral PhG may reflect the highly efficient process of memory encoding and extraction. Short-term experience may contribute to radiology interns spending less energy on employing memory resources when radiologists interpret the radiological images.

Limitation

It is important to note the limitations of our study. Firstly, the sample size was relatively small, which could limit the generalizability of the results. Future studies with larger sample sizes are needed to confirm the current findings. Secondly, the training duration for radiology interns was relatively short. Although the number of training cases for each participant was sufficient to acquire expertise, a longer training duration could potentially lead to different results. Therefore, future studies should consider longer training periods. Finally, a cross-sectional design was used in this study, which may limit the interpretation of the findings. Longitudinal studies are needed to better understand how visual experience affects brain dynamics in radiologists. Additionally, confounding factors such as long-term experience or congenital factors could have influenced the results. Therefore, future studies should consider controlling for these factors or using a longitudinal design to better understand the effects of visual experience on brain dynamics.

Conclusions

In conclusion, our findings suggest that visual experience can modulate the dynamics of the resting-state brain network, as reflected in multidimensional neurobehavioral components based on the expertise model of radiologists. These components are strongly interlinked with high-order cognitive and low-order visual factors, including attention control, memory, WM, decision making, and visual processing. These results provide a novel insight into the neural mechanism underlying visual expertise. Despite the limitations of our study, we believe that our findings contribute to the current understanding of how real-world visual experience affects brain activity and may have implications for radiologist training and clinical practice. Further research is needed to confirm and extend our findings.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by First Affiliated Hospital of Medical College in Xi'an Jiaotong University Subcommittee on Human Studies. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

Study design and interpretation of results: CJ and MD. Data collection or acquisition: HW, CJ, and MD. Statistical analysis: MD and XZ. Drafting the manuscript work: HW and RY. Essay revision: JW, CC, CJ, and MD. All authors read and approved the final version of the manuscript and agreed to publish.

Funding

This paper was supported by National Key R&D Program of China (Grant No. 2022YFF1202400), the National Natural Science Foundation of China (U19B2030), the Science and Technology Projects of Xi'an, China (No. 201809170CX11JC12), and the Fundamental Research Funds for the Central Universities (No. 20101236055).

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

References

Albert, N. B., Robertson, E. M., and Miall, R. C. (2009). The resting human brain and motor learning. *Curr. Biol.* 19, 1023–1027. doi: 10.1016/j.cub.2009.04.028

Annis, J., and Palmeri, T. J. (2019). Modeling memory dynamics in visual expertise. *J. Exp. Psychol. Learn. Mem. Cogn.* 45, 1599–1618. doi: 10.1037/xlm0000664

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Bi, T., Chen, J., Zhou, T., He, Y., and Fang, F. (2014). Function and structure of human left fusiform cortex are closely associated with perceptual learning of faces. *Curr. Biol.* 24, 222–227. doi: 10.1016/j.cub.2013.12.028

Bilalić, M. (2016). Revisiting the role of the fusiform face area in expertise. *J. Cogn. Neurosci.* 28, 1345–1357. doi: 10.1162/jocn_a_00974

Bilalic, M., Grottenthaler, T., Nägele, T., and Lindig, T. (2016). The faces in radiological images: fusiform face area supports radiological expertise. *Cereb. Cortex* 26, 1004–1014. doi: 10.1093/cercor/bhu272

Bilalić, M., Langner, R., Erb, M., and Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: a study with chess experts. *J. Exp. Psychol. Gen.* 139, 728–742. doi: 10.1037/a0020756

Briggs, R. G., Lin, Y. H., Dadario, N. B., Kim, S. J., Young, I. M., Bai, M. Y., et al. (2021). Anatomy and white matter connections of the middle frontal gyrus. *World Neurosurg.* 150, e520–e529. doi: 10.1016/j.wneu.2021.03.045

Canario, E., Chen, D., and Biswal, B. (2021). A review of resting-state fMRI and its use to examine psychiatric disorders. *Psychoradiology* 1, 42–53. doi: 10.1093/psyrad/kkab003

Cherkassky, V. (1997). The nature of statistical learning theory. *IEEE Trans. Neural Netw.* 8, 1564. doi: 10.1109/TNN.1997.641482

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M. A., et al. (2000). The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* 123 (Pt 2), 291–307. doi: 10.1093/brain/123.2.291

Devlin, J. T., Jamison, H. L., Gonnerman, L. M., and Matthews, P. M. (2006). The role of the posterior fusiform gyrus in reading. *J. Cogn. Neurosci.* 18, 911–922. doi: 10.1162/jocn.2006.18.6.911

Ding, X., Yang, Y., Stein, E. A., and Ross, T. J. (2015). Multivariate classification of smokers and nonsmokers using SVM-RFE on structural MRI images. *Hum. Brain Mapp.* 36, 4869–4879. doi: 10.1002/hbm.22956

Dong, M., Qin, W., Zhao, L., Yang, X., Yuan, K., Zeng, F., et al. (2014). Expertise modulates local regional homogeneity of spontaneous brain activity in the resting brain: an fMRI study using the model of skilled acupuncturists. *Hum. Brain Mapp.* 35, 1074–1084. doi: 10.1002/hbm.22235

Dong, M., Zhang, P., Chai, W., Zhang, X., Chen, B. T., Wang, H., et al. (2022). Early stage of radiological expertise modulates resting-state local coherence in the inferior temporal lobe. *Psychoradiology* 2, 199–206. doi: 10.1093/psyrad/kkac024

Donovan, T., and Litchfield, D. (2013). Looking for cancer: expertise related differences in searching and decision making. *Appl. Cogn. Psychol.* 27, 43–49. doi: 10.1002/acp.2869

Du, W., Cao, Z., Song, T., Li, Y., and Liang, Y. (2017). A feature selection method based on multiple kernel learning with expression profiles of different types. *BioData Min.* 10, 4. doi: 10.1186/s13040-017-0124-x

Engel, S. A., Harley, E. M., Pope, W. B., Villablanca, J. P., Mazziotta J. C., Enzmann, D., et al. (2009). "Activity in the fusiform face area supports expert perception in radiologists and does not depend upon holistic processing of images," in *Proceedings of the SPIE*, Volume 7263, eds B. Sahiner and D. J. Manning (SPIE). doi: 10.1117/12.812250

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., et al. (2016). The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb. Cortex* 26, 3508–3526. doi: 10.1093/cercor/bhw157

Gao, Y., Xiong, Z., Wang, X., Ren, H., Liu, R., Bai, B., et al. (2022). Abnormal degree centrality as a potential imaging biomarker for right temporal lobe epilepsy: a resting-state functional magnetic resonance imaging study and support vector machine analysis. *Neuroscience* 487, 198–206. doi: 10.1016/j.neuroscience.2022.02.004

Goold, J. E., and Meng, M. (2017). Categorical learning revealed in activity pattern of left fusiform cortex. *Hum. Brain Mapp.* 38, 3648–3658. doi: 10.1002/hbm.23620

Grill-Spector, K., Kourtzi, Z., and Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Res.* 41, 1409–1422. doi: 10.1016/S0042-6989(01)00073-6

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Haller, S., and Radue, E. W. (2005). What is different about a radiologist's brain? *Radiology* 236, 983–989. doi: 10.1148/radiol.2363041370

Harel, A. (2016). What is special about expertise? Visual expertise reveals the interactive nature of real-world object recognition. *Neuropsychologia* 83, 88–99. doi: 10.1016/j.neuropsychologia.2015.06.004

Harel, A., Kravitz, D., and Baker, C. I. (2013). Beyond perceptual expertise: revisiting the neural substrates of expert object recognition. *Front. Hum. Neurosci.* 7, 885. doi: 10.3389/fnhum.2013.00885

Harley, E. M., Pope, W. B., Villablanca, J. P., Mumford, J., Suh, R., Mazziotta, J. C., et al. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. *Cereb. Cortex* 19, 2746–2754. doi: 10.1093/cercor/bhp051

Hasler, B. P., Forbes, E. E., and Franzen, P. L. (2014). Time-of-day differences and short-term stability of the neural response to monetary reward: a pilot study. *Psychiatry Res.* 224, 22–27. doi: 10.1016/j.pscychresns.2014.07.005

Hendee, W. (2010). The handbook of medical image perception and techniques. *Med. Phys.* 37, 6112. doi: 10.1118/1.3505328

Japee, S., Holiday, K., Satyshur, M. D., Mukai, I., and Ungerleider, L. G. (2015). A role of right middle frontal gyrus in reorienting of attention: a case study. *Front. Syst. Neurosci.* 9, 23. doi: 10.3389/fnsys.2015.00023

Kanwisher, N., Woods, R. P., Iacoboni, M., and Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *J. Cogn. Neurosci.* 9, 133–142. doi: 10.1162/jocn.1997.9.1.133

Kesler, S. R., Lacayo, N. J., and Jo, B. (2011). A pilot study of an online cognitive rehabilitation program for executive function skills in children with cancer-related brain injury. *Brain Inj.* 25, 101–112. doi: 10.3109/02699052.2010.536194

Kim, Y. T., Seo, J. H., Song, H. J., Yoo, D. S., Lee, H. J., Lee, J., et al. (2011). Neural correlates related to action observation in expert archers. *Behav. Brain Res.* 223, 342–347. doi: 10.1016/j.bbr.2011.04.053

Kirk, U., Skov, M., Christensen, M. S., and Nygaard, N. (2009). Brain correlates of aesthetic expertise: a parametric fMRI study. *Brain Cogn.* 69, 306–315. doi: 10.1016/j.bandc.2008.08.004

Klingberg, T., O'Sullivan, B. T., and Roland, P. E. (1997). Bilateral activation of fronto-parietal networks by incrementing demand in a working memory task. *Cereb. Cortex* 7, 465–471. doi: 10.1093/cercor/7.5.465

Krawczyk, D. C., Boggan, A. L., McClelland, M. M., and Bartlett, J. C. (2011). The neural organization of perception in chess experts. *Neurosci. Lett.* 499, 64–69. doi: 10.1016/j.neulet.2011.05.033

Lee, D., Rushworth, M. F., Walton, M. E., Watanabe, M., and Sakagami, M. (2007). Functional specialization of the primate frontal cortex during decision making. *J. Neurosci.* 27, 8170–8173. doi: 10.1523/JNEUROSCI.1561-07.2007

Li, F., Huang, X., Tang, W., Yang, Y., Li, B., Kemp, G. J., et al. (2014). Multivariate pattern analysis of DTI reveals differential white matter in individuals with obsessive-compulsive disorder. *Hum. Brain Mapp.* 35, 2643–2651. doi: 10.1002/hbm.22357

Lin, Y. H., Dhanaraj, V., Mackenzie, A. E., Young, I. M., Tanglay, O., Briggs, R. G., et al. (2021). Anatomy and white matter connections of the parahippocampal gyrus. *World Neurosurg.* 148, e218–e226. doi: 10.1016/j.wneu.2020.12.136

Liu, D., Deng, J., Zhang, Z., Zhang, Z. Y., Sun, Y. G., Yang, T., et al. (2020). Orbitofrontal control of visual cortex gain promotes visual associative learning. *Nat. Commun.* 11, 2784. doi: 10.1038/s41467-020-16609-7

Liu, F., Wee, C. Y., Chen, H., and Shen, D. (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *Neuroimage* 84, 466–475. doi: 10.1016/j.neuroimage.2013.09.015

Liu, Y., and Lai, C. H. (2022). The alterations of degree centrality in the frontal lobe of patients with panic disorder. *Int. J. Med. Sci.* 19, 105–111. doi: 10.7150/ijms.65367

Martens, F., Bulthé, J., van Vliet, C., and Op de Beeck, H. (2018). Domain-general and domain-specific neural changes underlying visual expertise. *Neuroimage* 169, 80–93. doi: 10.1016/j.neuroimage.2017.12.013

McGugin, R. W., Van Gulick, A. E., Tamber-Rosenau, B. J., Ross, D. A., and Gauthier, I. (2015). Expertise effects in face-selective areas are robust to clutter and diverted attention, but not to competition. *Cereb. Cortex* 25, 2610–2622. doi: 10.1093/cercor/bhu060

Melo, M., Scarpin, D. J., Amaro, E. Jr., Passos, R. B., Sato, J. R., Friston, K. J., et al. (2011). How doctors generate diagnostic hypotheses: a study of radiological diagnosis with functional magnetic resonance imaging. *PLoS ONE* 6, e28752. doi: 10.1371/journal.pone.0028752

Miller, E. K., Lundqvist, M., and Bastos, A. M. (2018). Working memory 2.0. *Neuron* 100, 463–475. doi: 10.1016/j.neuron.2018.09.023

Moore, C. D., Cohen, M. X., and Ranganath, C. (2006). Neural mechanisms of expert skills in visual working memory. *J. Neurosci.* 26, 11187–11196. doi: 10.1523/JNEUROSCI.1873-06.2006

Muraskin, J., Sherwin, J., and Sajda, P. (2015). Knowing when not to swing: EEG evidence that enhanced perception-action coupling underlies baseball batter expertise. *Neuroimage* 123, 1–10. doi: 10.1016/j.neuroimage.2015.08.028

Nakashima, R., Watanabe, C., Maeda, E., Yoshikawa, T., Matsuda, I., Miki, S., et al. (2015). The effect of expert knowledge on medical search: medical experts have specialized abilities for detecting serious lesions. *Psychol. Res.* 79, 729–738. doi: 10.1007/s00426-014-0616-y

Ouellette, D. J., Van Staalduinen, E., Hussaini, S. H., Govindarajan, S. T., Stefancin, P., Hsu, D. L., et al. (2020). Functional, anatomical and diffusion tensor MRI study of radiology expertise. *PLoS ONE* 15, e0231900. doi: 10.1371/journal.pone.0231900

Peelen, M. V., and Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* 93, 603–608. doi: 10.1152/jn.00513.2004

Rasmussen, P. M., Madsen, K. H., Lund, T. E., and Hansen, L. K. (2011). Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *Neuroimage* 55, 1120–1131. doi: 10.1016/j.neuroimage.2010.12.035

Reynolds, B. B., Stanton, A. N., Soldozy, S., Goodkin, H. P., Wintermark, M., and Druzgal, T. J. (2018). Investigating the effects of subconcussion on functional connectivity using mass-univariate and multivariate approaches. *Brain Imaging Behav.* 12, 1332–1345. doi: 10.1007/s11682-017-9790-z

Rudebeck, P. H., and Rich, E. L. (2018). Orbitofrontal cortex. *Curr. Biol.* 28, R1083–r1088. doi: 10.1016/j.cub.2018.07.018

Smitha, K. A., Akhil Raja, K., Arun, K. M., Rajesh, P. G., Thomas, B., Kapilamoorthy, T. R., et al. (2017). Resting state fMRI: a review on methods in resting state connectivity analysis and resting state networks. *Neuroradiol. J.* 30, 305–317. doi: 10.1177/1971400917697342

Spiers, H. J., and Maguire, E. A. (2006). Thoughts, behaviour, and brain dynamics during navigation in the real world. *Neuroimage* 31, 1826–1840. doi: 10.1016/j.neuroimage.2006.01.037

Su, J., Zhang, X., Zhang, Z., Wang, H., Wu, J., Shi, G., et al. (2022). Real-world visual experience alters baseline brain activity in the resting state: a longitudinal study using expertise model of radiologists. *Front. Neurosci.* 16, 904623. doi: 10.3389/fnins.2022.904623

van der Linden, M., Murre, J. M., and van Turennout, M. (2008). Birds of a feather flock together: experience-driven formation of visual object categories in human ventral temporal cortex. *PLoS ONE* 3, e3995. doi: 10.1371/journal.pone.0003995

Volz, K. G., Schubotz, R. I., and von Cramon, D. Y. (2006). Decision-making and the frontal lobes. *Curr. Opin. Neurol.* 19, 401–406. doi: 10.1097/01.wco.0000236621.83872.71

Wang, Y., Jin, C., Yin, Z., Wang, H., Ji, M., Dong, M., et al. (2021). Visual experience modulates whole-brain connectivity dynamics: a resting-state fMRI study using the model of radiologists. *Hum. Brain Mapp.* 42, 4538–4554. doi: 10.1002/hbm.25563

Weiner, K. S., and Zilles, K. (2016). The anatomical and functional specialization of the fusiform gyrus. *Neuropsychologia* 83, 48–62. doi: 10.1016/j.neuropsychologia.2015.06.033

Westerberg, H., and Klingberg, T. (2007). Changes in cortical activity after training of working memory–a single-subject analysis. *Physiol. Behav.* 92, 186–192. doi: 10.1016/j.physbeh.2007.05.041

Xu, Y. (2005). Revisiting the role of the fusiform face area in visual expertise. *Cereb. Cortex* 15, 1234–1242. doi: 10.1093/cercor/bhi006

Zhang, T., Dong, M., Wang, H., Jia, R., Li, F., Ni, X., et al. (2022). Visual expertise modulates baseline brain activity: a preliminary resting-state fMRI study using expertise model of radiologists. *BMC Neurosci.* 23, 24. doi: 10.1186/s12868-022-00707-x

Zuo, X. N., and Xing, X. X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neurosci. Biobehav. Rev.* 45, 100–118. doi: 10.1016/j.neubiorev.2014.05.009

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 02 June 2023
DOI 10.3389/fnins.2023.1152222

OPEN ACCESS

EDITED BY
Dingwen Zhang,
Northwestern Polytechnic University,
United States

REVIEWED BY
Mingming Gong,
The University of Melbourne, Australia
Dongjin Song,
University of Connecticut, United States

*CORRESPONDENCE
Zhiwei Ye
✉ hgcsyzw@hbut.edu.cn

RECEIVED 27 January 2023
ACCEPTED 17 April 2023
PUBLISHED 02 June 2023

CITATION
Mei M, Ye Z and Cha Y (2023) An integrated convolutional neural network for classifying small pulmonary solid nodules.
Front. Neurosci. 17:1152222.
doi: 10.3389/fnins.2023.1152222

An integrated convolutional neural network for classifying small pulmonary solid nodules

Mengqing Mei[1], Zhiwei Ye[1]* and Yunfei Zha[2]

[1]School of Computer Science, Hubei University of Technology, Wuhan, China, [2]Department of Radiology, Renmin Hospital of Wuhan University, Wuhan, China

Achieving accurate classification of benign and malignant pulmonary nodules is essential for treating some diseases. However, traditional typing methods have difficulty obtaining satisfactory results on small pulmonary solid nodules, mainly caused by two aspects: (1) noise interference from other tissue information; (2) missing features of small nodules caused by downsampling in traditional convolutional neural networks. To solve these problems, this paper proposes a new typing method to improve the diagnosis rate of small pulmonary solid nodules in CT images. Specifically, first, we introduce the Otsu thresholding algorithm to preprocess the data and filter the interference information. Then, to acquire more small nodule features, we add parallel radiomics to the 3D convolutional neural network. Radiomics can extract a large number of quantitative features from medical images. Finally, the classifier generated more accurate results by the visual and radiomic features. In the experiments, we tested the proposed method on multiple data sets, and the proposed method outperformed other methods in the small pulmonary solid nodule classification task. In addition, various groups of ablation experiments demonstrated that the Otsu thresholding algorithm and radiomics are helpful for the judgment of small nodules and proved that the Otsu thresholding algorithm is more flexible than the manual thresholding algorithm.

KEYWORDS

medical image analysis, neural networks, classification, pulmonary solid nodules, feature extraction

1. Introduction

Pulmonary nodule classification is an important task that judges the benignity and malignancy of pulmonary nodules by computer techniques. Deep learning methods based on convolutional neural networks are the most common methods for pulmonary nodule classification, which can be divided into 2D CNN based methods (Setio et al., 2016; Sori et al., 2019) and 3D CNN based methods (Shi et al., 2021; Tsai and Peng, 2022). In general, the 2D CNN based methods for pulmonary nodule classification have three steps: first, the 3D CT images are sliced; then, the features are extracted by 2D CNN; finally, the extracted features are input to the classifier to obtain the results. Shen et al. (2017) constructed an end-to-end architecture, which used neural networks for feature extraction instead of complex nodule segmentation and manual fabrication, and achieved better classification accuracy.

The pulmonary nodule images are three-dimensional, and extracting two-dimensional slices with only one view of the image could easily cause information deficiency. Therefore, some studies have proposed multi-view slicing methods. Xie et al. (2018) proposed to learn the features of 3D pulmonary nodules by 9 fixed views, different views are learned using

different sub-models. Al-Shabi et al. (2019) proposed a local-global neural network, which uses the residual module with the convolution kernel size of 3×3 to extract local features, and global features are extracted by self-attention layers, the combination of both features achieves better classification results. Considering the three-dimensional characteristics of 3D CT images, many works based on 3D CNNs have appeared in recent years. Kang et al. (2017) proposed a 3D multi-view convolutional neural network to better utilize 3D contextual information and extracted more discriminative features. Liu et al. (2021) proposed a contextual attention network (CA-Met), CA-Met extracts the nodules and surrounding tissues features by contextual attention and then fuses the two features into the classifier for prediction. Zhang et al. (2018) introduced 3D dilated convolution into the base model, which helps the model retain more image information and acquire multi-scale features, leading to more accurate nodule classification results. Huang et al. (2022) proposed a novel neural network based on self-supervised learning. This network learns labeled and unlabeled data to overcome the problem of insufficient labeled samples and eliminates noisy information interference by data preprocessing. Although 3D CNN achieves great classification results, it faces the problems of large computation and complex network structure. Therefore, the pulmonary nodule classification still needs further research to play a greater role in the medical career.

2. A 3D residual convolutional neural network

Providing accurate classification of small pulmonary solid nodules in CT images is significant. For this reason, this paper proposes a classification method for small pulmonary solid nodules. Specifically, we use the Otsu thresholding algorithm to reduce the noise interference from pulmonary tissues around the nodules. The proposed method extracts two modal features through a 3D residual convolutional neural network and radiomics. Then we fuse the visual and radiomic features into the classifier, which helps the model better predict benign and malignant nodules. As shown in Figure 1, the complete framework of the classification method consists of three parts: (1) data preprocessing based on the Otsu thresholding algorithm; (2) the interactive learning based on two modal features of three-dimensional residual convolutional neural network and radiomics; (3) small solid nodules classification.

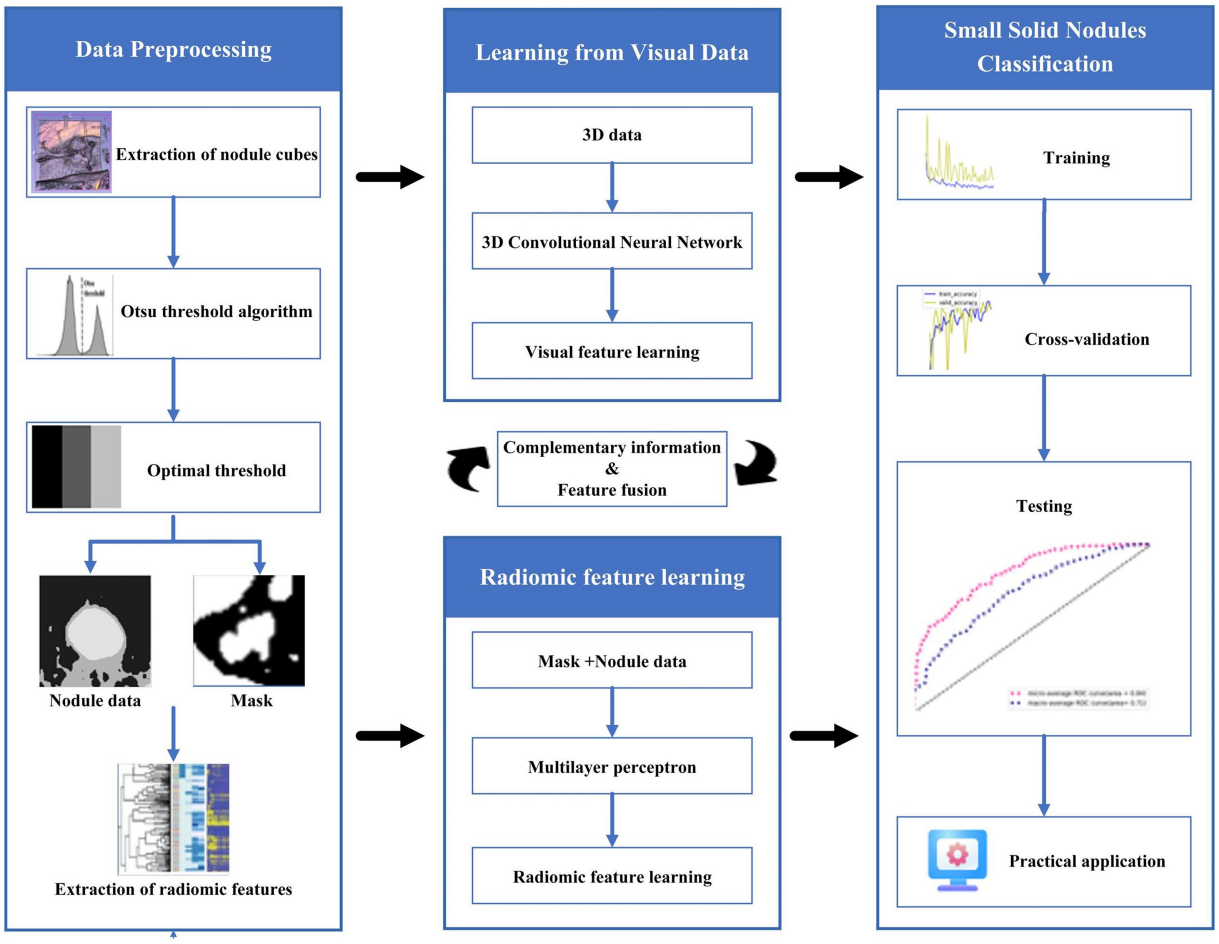


FIGURE 1
Flowchart of the proposed algorithm.

10.3389/fnins.2023.1152222

2.1. Data preprocessing

This paper uses the Otsu threshold algorithm for multi-threshold segmentation preprocessing of medical CT images. This algorithm divides the images into three classes, C1, C2, and C3, by two thresholds, K1 and K2, and then obtains the optimal threshold by the maximum inter-class variance. The inter-class variance is formulated as follows:

$$\sigma^2 = \sum_{k=1}^{3} P_k \left(m_k - m_G\right)^2 \tag{1}$$

where k is the corresponding class. P_k denotes the probability of class k, m refers to the mean gray value of the kth class and the formula is as follows:

$$P_k = \sum_{i \in C_k} p_i \tag{2}$$

where p_i is the probability of the gray level being i. The probability of different classes can be written as:

$$P_1 = \sum_{i=0}^{K_1} p_i, P_2 = \sum_{i=K_1+1}^{K_2} p_i, P_3 = \sum_{i=K_2+1}^{255} p_i \tag{3}$$

The mean gray value of the kth class can be defined as:

$$m_k = \frac{1}{P_k} \sum_{i \in C_k} ip_i \tag{4}$$

The mean gray value of three classes can be formulated as:

$$m_1 = \frac{1}{P_1} \sum_{i=0}^{K_1} ip_i,\ m_2 = \frac{1}{P_2} \sum_{i=K_1+1}^{K_2} ip_i,\ m_3 = \frac{1}{P_3} \sum_{i=K_2+1}^{255} ip_i \tag{5}$$

The mean gray value of the entire image is calculated as:

$$m_G = \sum_{i=0}^{255} ip_i \tag{6}$$

Finally, two optimal thresholds K_1^* and K_2^* are obtained by maximizing $\sigma^2\left(K_1, K_2\right)$, the formula is as follows:

$$\sigma^2\left(K_1^*, K_2^*\right) = \max_{0<K_1<K_2<255} \sigma^2\left(K_1, K_2\right) \tag{7}$$

If the maximum value of inter-class variance is not unique, the corresponding optimal thresholds K_1^* and K_2^* are averaged to obtain the final threshold. The Otsu algorithm is described as follows:

Algorithm: Otsu algorithm

Input: grayscale images generated from medical CT images

Output: optimal thresholds K_1^* and K_2^*

calculate the normalized histogram and the probability p_i that the gray level is i ($i = 0, 1, \ldots, 255$) from the input image;

calculate the mean gray value $m_G = \sum_{i=0}^{255} ip_i$ of the entire image;

For $K1 = 1{:}253$

For $K2 = K1 + 1{:}254$

calculate the probability of three classes $P_1 = \sum_{i=0}^{K_1} p_i, P_2 = \sum_{i=K_1+1}^{K_2} p_i, P_3 = \sum_{i=K_2+1}^{255} p_i$;

calculate the mean gray value $m_1 = \frac{1}{P_1} \sum_{i=0}^{K_1} ip_i, m_2 = \frac{1}{P_2} \sum_{i=K_1+1}^{K_2} ip_i, m_3 = \frac{1}{P_3} \sum_{i=K_2+1}^{255} ip_i$;

calculate the inter-class variance $\sigma^2(K_1, K_2) = \sum_{k=1}^{3} P_k (m_k - m_G)^2$;

End

End

obtain two optimal thresholds K_1^* and K_2^* by (7);

If $size\left(K_1^*, 1\right) > 1$

$K_1^* = mean\left(K_1^*\right)$;

End

If $size\left(K_2^*, 1\right) > 1$

$K_2^* = mean\left(K_2^*\right)$;

End

The bounding boxes of nodular lesions are mainly labeled according to the nodule location information provided by professional radiologists, and the examples are shown in Figure 2. The labeled information is a rectangle composed of six-coordinate points in three-dimensional space, and the six-coordinate points are denoted as $x_{max}?x_{min}?\ y_{max}?y_{min}?z_{max}?z_{min}$ Due to the input requirement of the convolutional neural network is a cube, and the resampling will change the nodule shape, this paper selects the maximum value of H, W, and D as the side length of the cube and uses the padding operation to ensure the integrity of the nodule information, the pairs in three-dimensional space are shown in Figure 3. In addition, considering the interference factors such as blood vessels, air bubbles, and lung lobes, this paper extracts the nodule proper and its edge within a smaller error range by the Otsu thresholding algorithm. Specifically, a three-band thresholding classification method is used in this section, and the last two bands are selected as the retained information. As shown in Figure 4, the black areas indicate the parts that will be ignored, and the gray and white areas are the information that will be retained.

In common medical image processing, three-dimensional data are treated as N two-dimensional slices, computed by two-dimensional convolution kernels to obtain feature maps. These methods suffer from the problem of missing correlations between different slices, leading to inconsistency in the conclusions. In the small solid pulmonary nodule classification task, the shape of the same small pulmonary solid nodule varies

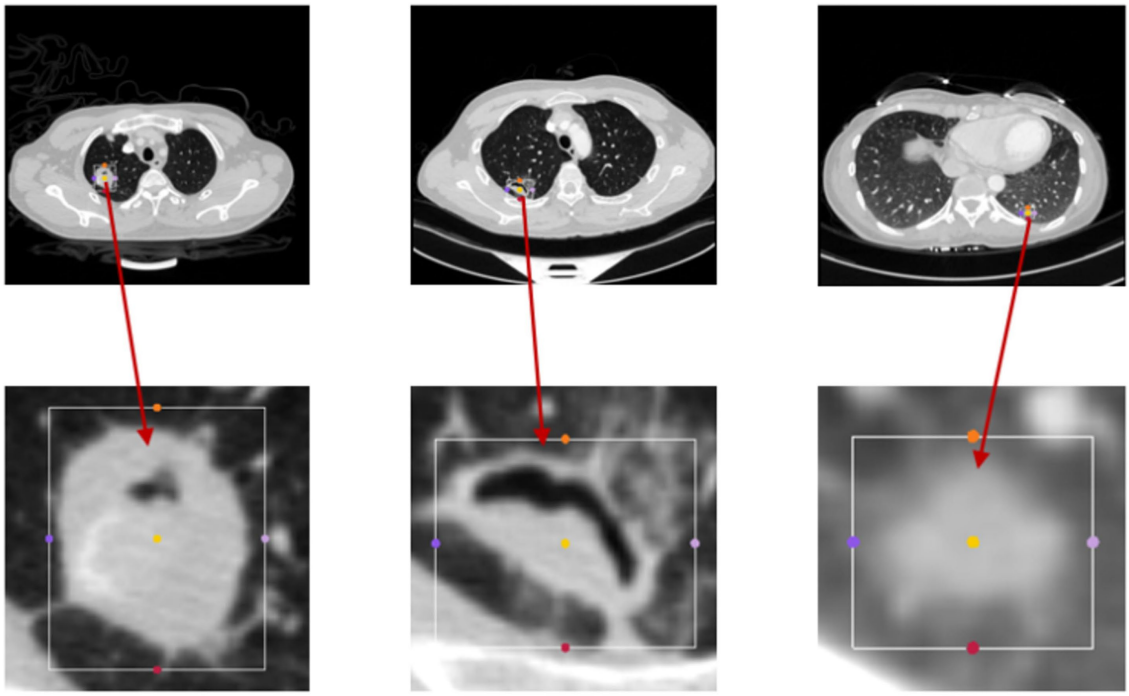


FIGURE 2
The figure displays the bounding boxes of nodular lesions.

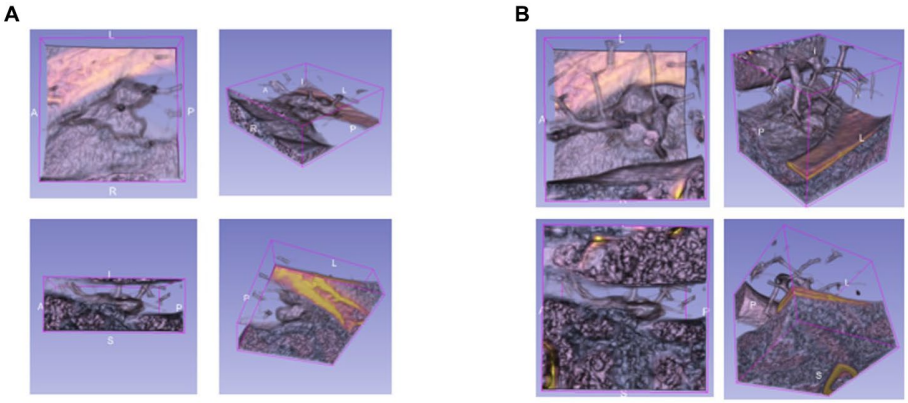


FIGURE 3
The example of the pulmonary nodules in three dimensions, **(A)** rectangular, **(B)** cubic.

greatly between different slices, so it is not easy to achieve the expected results by training with the traditional scheme. Therefore, this paper utilizes the cube containing the whole nodule information for training. The examples of two-dimensional slice information are shown in Figure 5.

This paper uses an algorithm based on PyRadiomics (a radiomic features extraction package developed by the Harvard Medical School team) to extract radiomic features of nodules. The various features extracted include the following classes: First Order Statistics has 19 features, which mainly include the magnitude of voxel values in the cube, the maximum, minimum and the range of the grayscale value of the lesion area; Shaped-Based (3D), which includes 16 features, such as the voxel volume, surface area, sphericity and surface area to volume ratio; Shaped-Based (2D) have 10 features, mainly including mesh surface, pixel surface, and perimeter to surface ratio; Gray Level Co-occurence Matrix have 24 features, mainly including autocorrelation, joint average, and cluster shade; gray level run length matrix have 16 features, mainly including short and long run emphasis, gray level non-uniformity and run length non-uniformity; gray level size zone matrix have 16 features, especially including small and large area emphasis; Neighbouring Gray Tone Difference Matrix have 16 features, including coarseness, contrast, busyness, complexity and strength; Gray

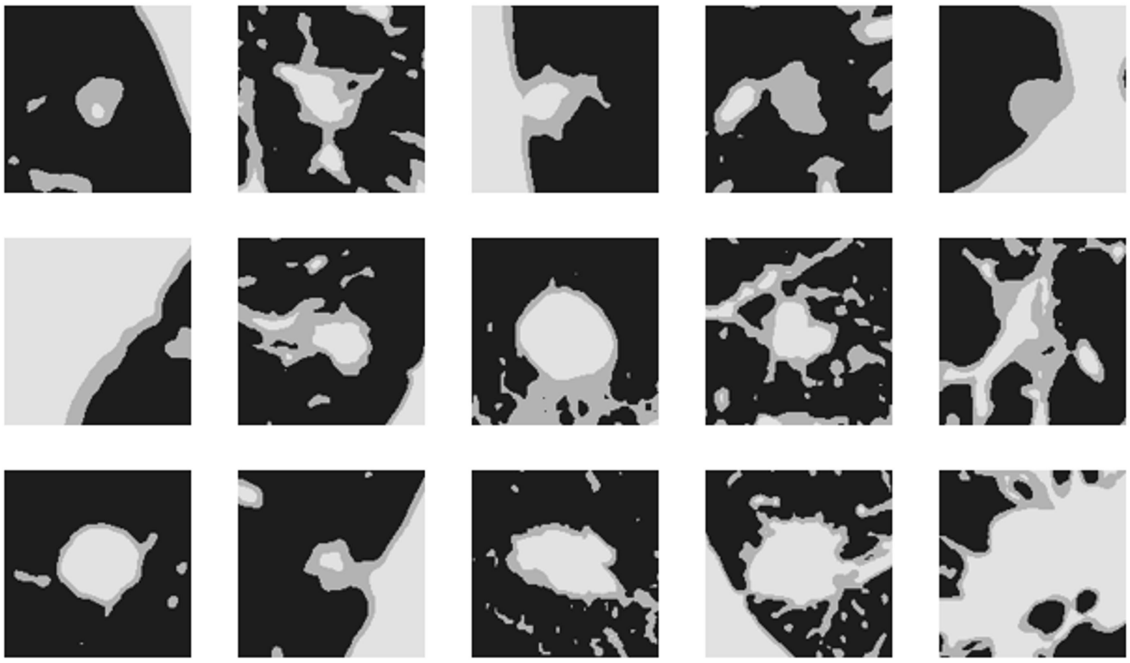


FIGURE 4
The results of the Otsu thresholding algorithm on CT data.

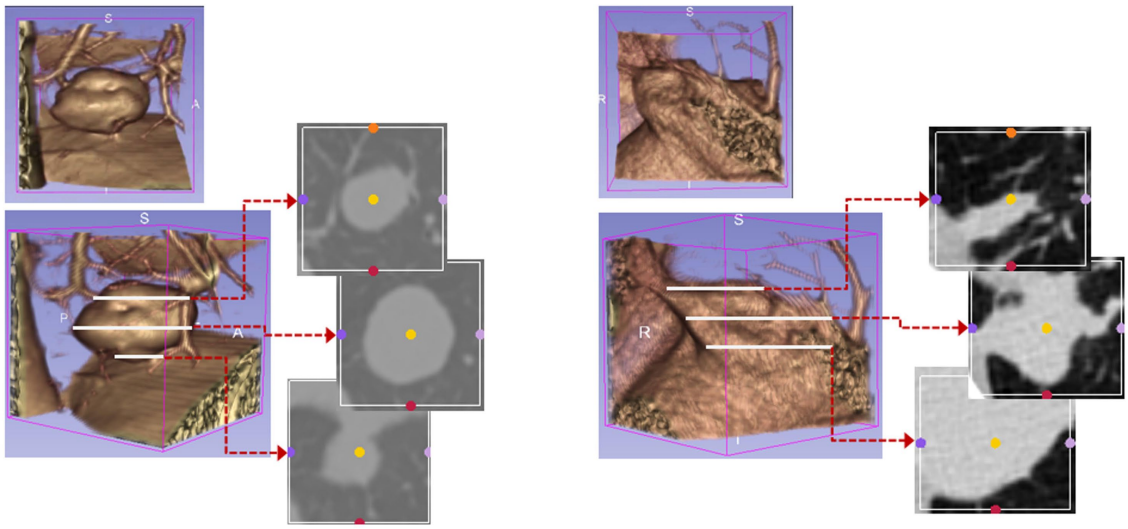


FIGURE 5
The 3D information of small pulmonary nodules and their corresponding 2D slices.

Level Dependence Matrix have 14 features, mainly including small and large dependence emphasis, gray level variance, dependence entropy. In addition, this paper uses the results from the Otsu algorithm as the input mask to ensure that the information obtained from radiomics is accurate, and the extraction process is shown in Figure 6.

Few images and unbalanced class distribution are common problems in medical image processing. This paper uses a data enhancement algorithm based on MONAI to alleviate these problems, which can strengthen the neural network's generalization. The data enhancement approaches used for the nodular cube include random 3D image rotation, random 3D image flip, and random affine transformation; those approaches will not act on the radiomic features. Considering the non-uniform classification of benign and malignant nodules in the training dataset, the lesser class (benign) will get more enhancement. The nodules will randomly flip along an axis with the probability of 70% benign and 40% malignant. Then, they will be affine transformed with the

probability of 70% benign and 40% malignant. Finally, they are randomly rotated with the same probability (the maximum rotation angle for benign is 35° and for malignant is 30°).

2.2. Fusion learning of 3D residual convolutional neural networks and radiomics

To better utilize the information of small pulmonary solid nodules in three-dimensional space, this paper combines 3D residual convolutional neural networks and radiomics to achieve nodule classification using complementary information between different modalities. As shown in Figure 7, the 3D visual and radiomic features information of small pulmonary solid nodules are input to 3D residual convolutional neural networks and multilayer perceptron, respectively. The two features are fused, and the results are output through the prediction layer.

In the training process, the backpropagation of gradients is usually affected by the depth of the network, and more layers easily cause poor training results. He et al. introduced skip connections on convolutional neural networks to alleviate the gradient disappearance due to the over-deepening of the network. Inspired by this structure, this paper adds a skip connection structure to the 3D convolutional module, and the 3D skip connection module is shown in Figure 8. The equation of this module can be defined as:

$$R(x) = W_l\left(f\left(W_{l-1}(x)\right)\right) \tag{8}$$

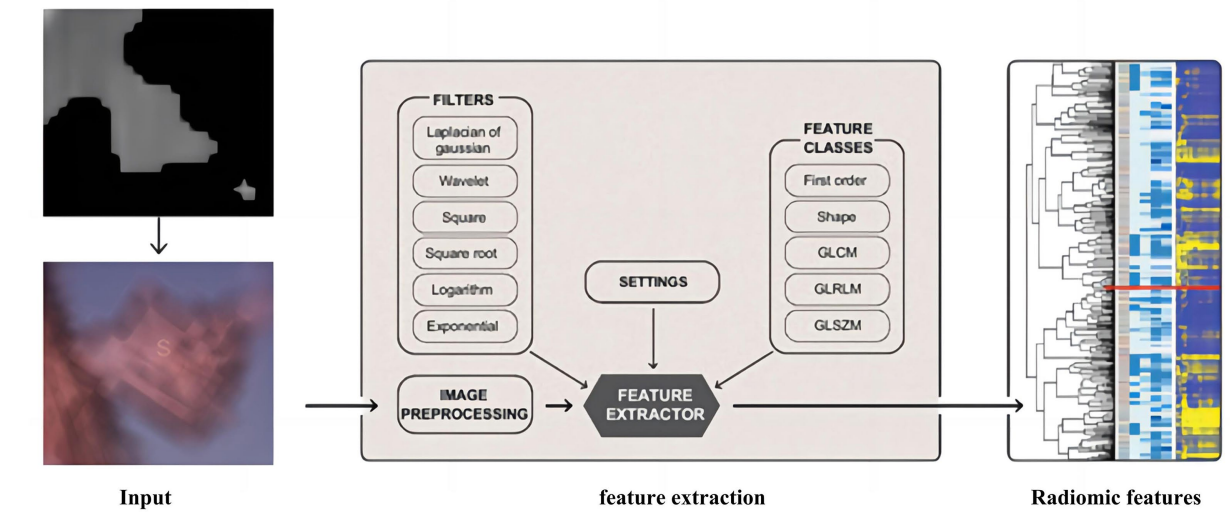


FIGURE 6
The extraction process of radiomics.

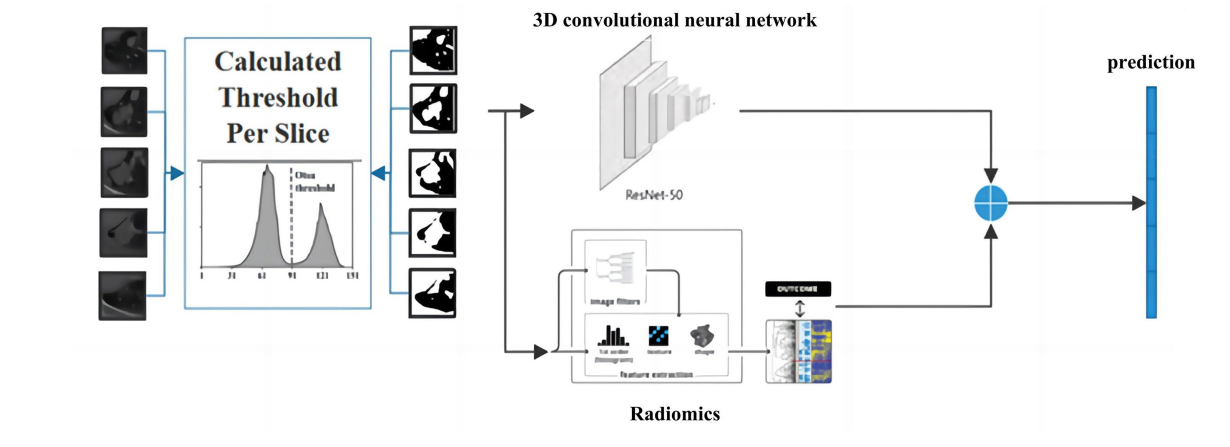


FIGURE 7
The architecture of the proposed model.

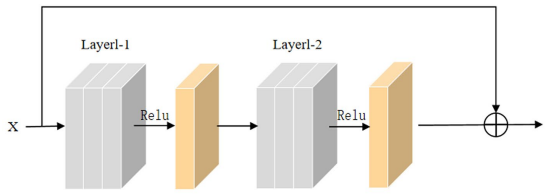


FIGURE 8
Diagram of skip connection module.

$$U(x) = R(x) + M(x) \tag{9}$$

where x is the input to the skip connection module, $M(x)$ is the skip connection, $U(x)$ is the original function, and $R(x)$ is the residual function.

In the 3D convolution operation, the input is convolved by the 3D convolution kernel, and bias is added. The 3D feature map is output using the normalization layer and the nonlinear activation unit. The formula for 3D convolution operation is as follows:

$$\text{out}\left(N_i, C_{\text{out}_j}\right) = \text{bias}\left(C_{\text{out}_j} + \sum_{k=0}^{C_{\text{in}}-1} \text{weight}\left(C_{\text{out}_j}, k\right) * \text{input}\left(N_i, k\right)\right) \tag{10}$$

where N, C, D, H, and W represent the batch size, number of channels, number of slices, length of slices, and width of slices, respectively. The operators ? are 3D interpolation operations. Meanwhile, Rectified Linear Unit (ReLU) and 3D BatchNorm are used in the model.

The residual convolutional neural network is designed based on the traditional convolutional neural network, which uses multiply subsampled to expand the local receptive field. However, the multiple subsampled operations will lead to a severe loss of small nodule features in the small solid pulmonary nodule classification task. This paper uses radiomics to solve the problem of information loss caused by subsampled. Radiomic features have correlation and complementarity with visual features. Specifically, the fusion of radiomic and visual features can complement the lost information from the statistical dimension of shape representation, making the network more robust. In addition, this paper uses dimensionality reduction to keep the visual features and radiomic features in the same dimension, and the two features are fused and input to the classifier to generate predictions. The detailed structure of the proposed network model is shown in Table 1.

TABLE 1 The overall structure of the proposed model.

Layer name	Output size	3D-ResNET50	Radiomics
Conv_1	128×128×128	7×7, 64, stride 2	/
		3×3 max pool, stride 2	/
Layer_1	56×56×56	$\begin{bmatrix} 1x1x1,64 \\ 3x3x3,64 \\ 1x1x1,256 \end{bmatrix} x3$	/
Layer_2	28×28×28	$\begin{bmatrix} 1x1x1,128 \\ 3x3x3,128 \\ 1x1x1,512 \end{bmatrix} x4$	/
Layer_3	14×14×14	$\begin{bmatrix} 1x1x1,256 \\ 3x3x3,256 \\ 1x1x1,1024 \end{bmatrix} x6$	
Layer_4	7×7×7	$\begin{bmatrix} 1x1x1,512 \\ 3x3x3,512 \\ 1x1x1,2048 \end{bmatrix} x3$	
	1×1×1	Average pool	107-d
Layer_Linear		Linear(256-d)	
Layer_MLP			MLP (256-d)
Layer_concat	512-d	Concat (Layer_Linear, Layer_MLP)	

2.3. Training and prediction

The training of the model can be divided into three steps. First, we follow the steps in the previous section to extract the cube containing the nodule area and obtain the mask of the nodule tissue by the Otsu thresholding algorithm. The mask can filter out the useless regions in the cube and also serve as the annotation for the radiomics extraction; Second, we resample the processed cube and expand the training data by the data enhancement methods, such as random flip and random radiation; Third, the enhanced cubes and the corresponding radiomics are input into the network for training.

After training, the data in the test set can perform nodule classification with the saved weights. We first extract the cube containing the nodules and obtain the optimal threshold by the Otsu thresholding algorithm. Then the extracted nodules are processed as in the training stage to obtain the radiomic features and the filtered nodule cube. Finally, the two groups of features are input to the trained network model to get the results.

3. Experimental results and analysis

The Adam optimizer trains the model with momentum set to $\text{beta}_1 = 0.9, \text{beta}_2 = 0.999$. We train 100 iterations, with the initial learning rate set to 0.001 and the learning rate dropping by 10% every ten iterations. In addition, dropout is set to 0.5, the batch size is set to 16, and the loss function is binary cross-entropy.

The datasets in the experiments came from the cooperative hospitals, which are non-public datasets, and the training set and the validation set have 1,429 samples in total. These samples are randomly divided into five subsets, represented as $\text{subset}_0, \text{subset}_1, \text{subset}_2, \text{subset}_3, \text{subset}_4$. One of them is taken as the validation set for each experiment, while the other subsets are

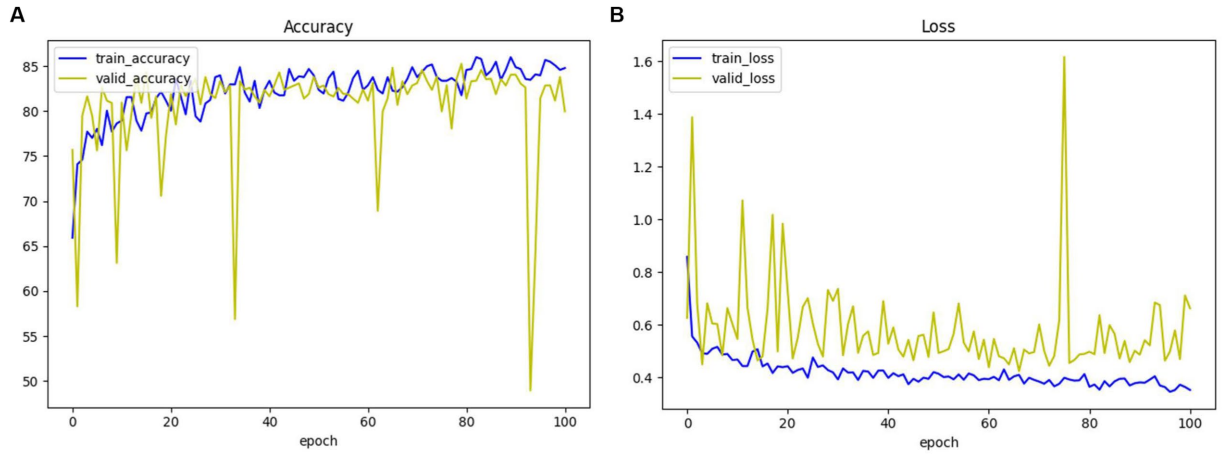


FIGURE 9
The accuracy and loss of 3D ResNet+Radiomics network in training and validation. **(A)** Loss. **(B)** Accuracy.

used as the training set. In addition, 200 additional samples were collected as the test set to verify the robustness of the model and to ensure its effectiveness in practice, and these data were collected from different hospitals in independent time and independent devices.

The experiments mainly use classification accuracy, receiver operating characteristic (ROC) curve, and area under curve (AUC) values to analyze and evaluate the results. The accuracy is an indicator that can directly judge the inference ability of the model and can be defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP, TN, FN, and FP are the numbers of true positive, true negative, false negative, and false positive pixels, respectively. Considering the influence of the optimal threshold in the task of small pulmonary solid nodules, we introduced ROC into the evaluation index and plotted ROC curves based on the false positive rate (FPR) and true positive rate (TPR) of the predicted results.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, (0 \le \text{TPR} \le 1).$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, (0 \le \text{FPR} \le 1).$$

AUC is the area covered by the ROC curve, and the value can visually reflect the good or bad performance of the classifier under different thresholds, which range from 0 to 1. The higher the ACU value, the better the classifier performance.

For the convenience of expressing the model, we will use 3D ResNet to denote the three-dimensional residual convolutional neural network, 3D VGG to denote the three-dimensional VGG model, and 3D ResNext to represent the three-dimensional ResNeXt model in the following. The accuracy and loss of the 3D ResNet + Radiomics network in training and validation are presented in Figure 9, and these values can be used as a basis for model convergence when they tend to be stable.

The same model may find different local optimal solutions during two training processes, which causes differences in model performance. This paper selects the model weights that perform best on the validation set for testing. As shown in Figure 10, the same model under different training achieves different results on the validation set, and it can be observed that the model performance of result 2 is better than result 1.

Ablation experiments of radiomics

To verify the effectiveness of radiomics for the classification task, we compared five different models, including 3D ResNet, 3D ResNeXt, 3D VGG, 3D DenseNet, and the proposed 3D ResNet + Radiomics model, and these experiments used the same preprocessing. Figure 11 reports the AUC/ROC of different models on the same test set, and the proposed model obtained better results. Radiomics alleviate the effect of information loss caused by downsampling in the convolutional neural network, which enables the model to learn more features and show more robustness on the test set.

Comparison experiments between the Otsu thresholding algorithm and manual thresholding algorithm

To verify the superiority of the Otsu thresholding algorithm for preprocessing, we compared the Otsu thresholding algorithm and the manual thresholding algorithm. Figure 12

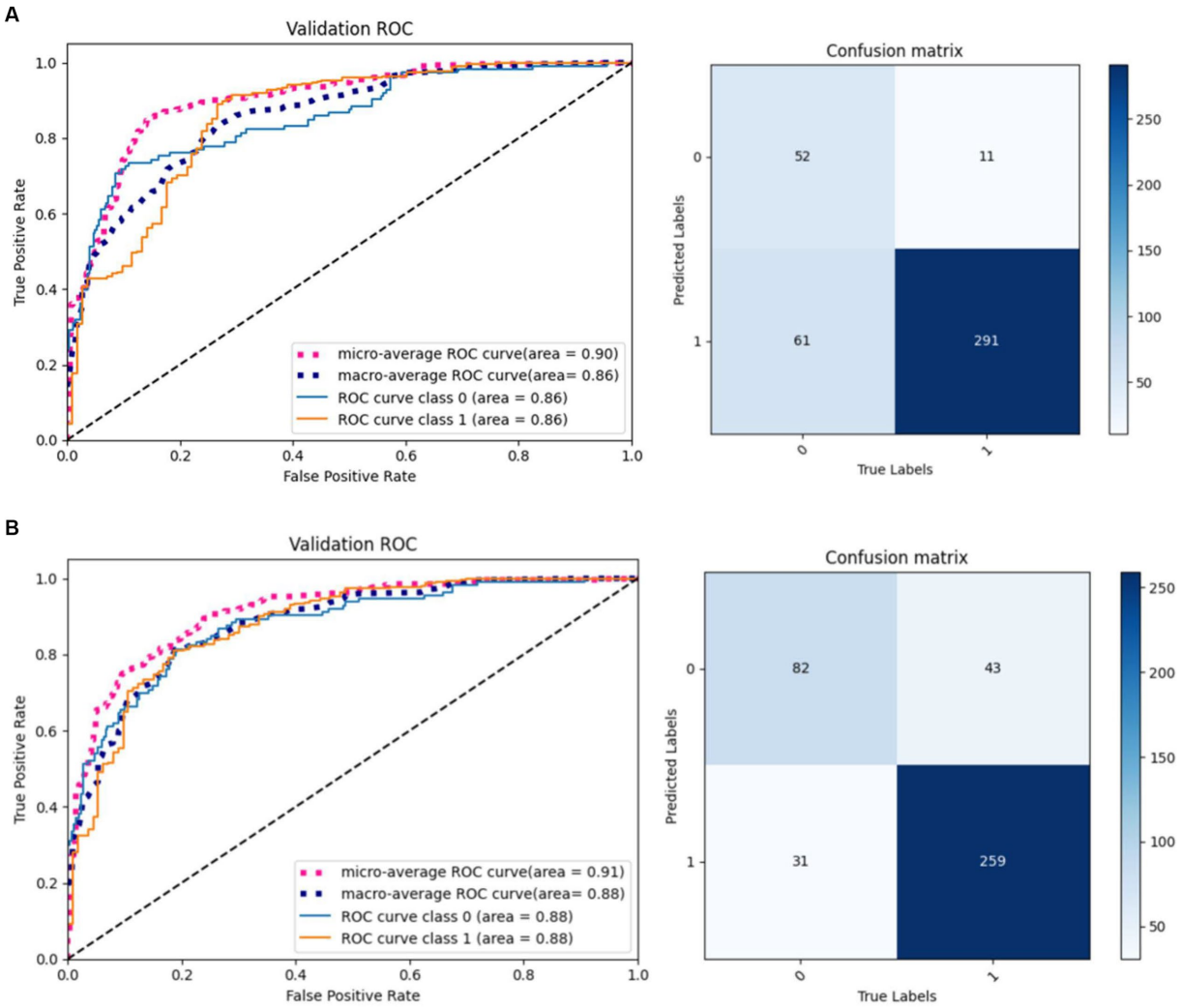


FIGURE 10
Results of the proposed model in two training. **(A)** Result 1. **(B)** Result 2.

TABLE 2 Comparison between different methods with and without our refinement.

Method	M	O	R	Micro-Average AUC	Macro-Average AUC
3D Resnet50	×	√	×	85%	67%
3D ResNext	×	√	×	76%	68%
3D DenseNet	×	√	×	72%	74%
3D VGG11	×	√	×	61%	68%
3D Resne50 + Radiomics	√	×	√	83%	64%
3D Resne50 + Radiomics (ours)	×	√	√	84%	71%

shows the visualization results of both algorithms on the same pulmonary nodules. The manual thresholding algorithm sets the range of HU values of CT images from 0 to 300, which is referenced to the common range of human tissues and experimental results and is more sensitive to solid tissues and lung cavities.

The experimental results of the two algorithms are shown in Figure 13. The Otsu thresholding algorithm filters the tissue information around the small solid pulmonary nodules and achieves better results. On the one hand, the Otsu thresholding algorithm can provide flexible threshold adjustment to prevent filtering out the valuable information of nodules. On the other

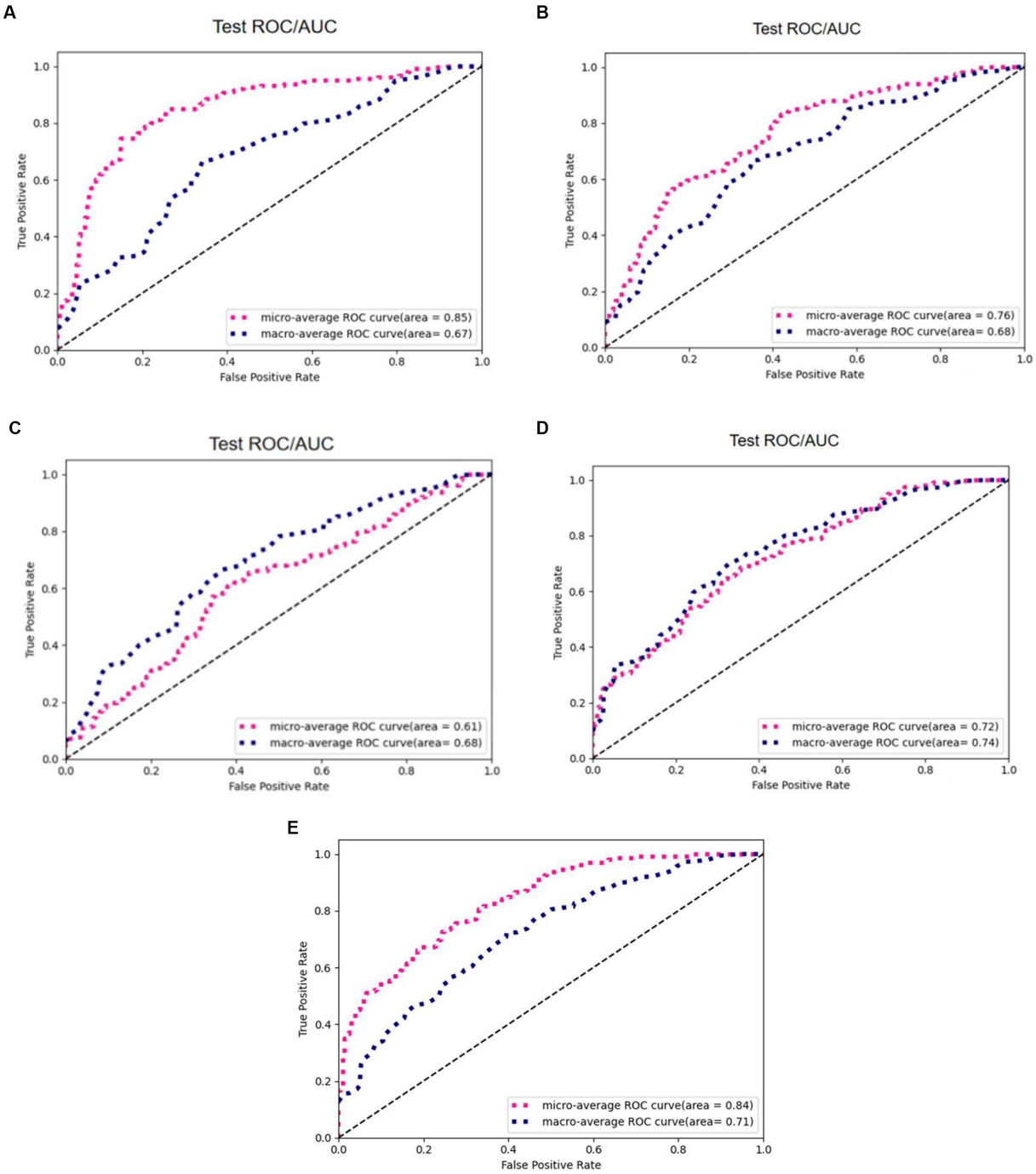


FIGURE 11
AUC/ROC of the models on the test set. **(A)** 3D ResNet50. **(B)** 3D ResNext. **(C)** 3D VGG. **(D)** 3D DenseNet. **(E)** 3D ResNet50+Radiomics.

hand, the manually delineated thresholds lack the necessary flexibility in practical application and perform unstably on CT data facing different devices and batches, making the process of filtering information risky and unsuitable for application in practice.

Table 2 reports the results of multiple algorithms under different conditions, and the proposed algorithm achieves the best results. Compared with 3D ResNet50, 3D ResNet50 + Radiomics has a more significant improvement in Macro-Average AUC while keeping Micro-Average AUC unchanged.

4. Conclusion

In this work, we propose a typing method based on the Otsu thresholding algorithm for small pulmonary solid nodules. The Otsu

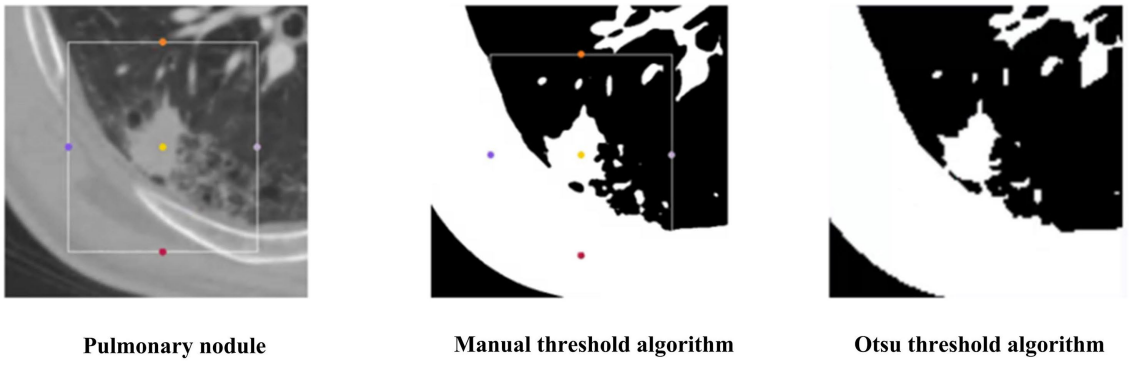


FIGURE 12
The segmentation results of two threshold algorithms.

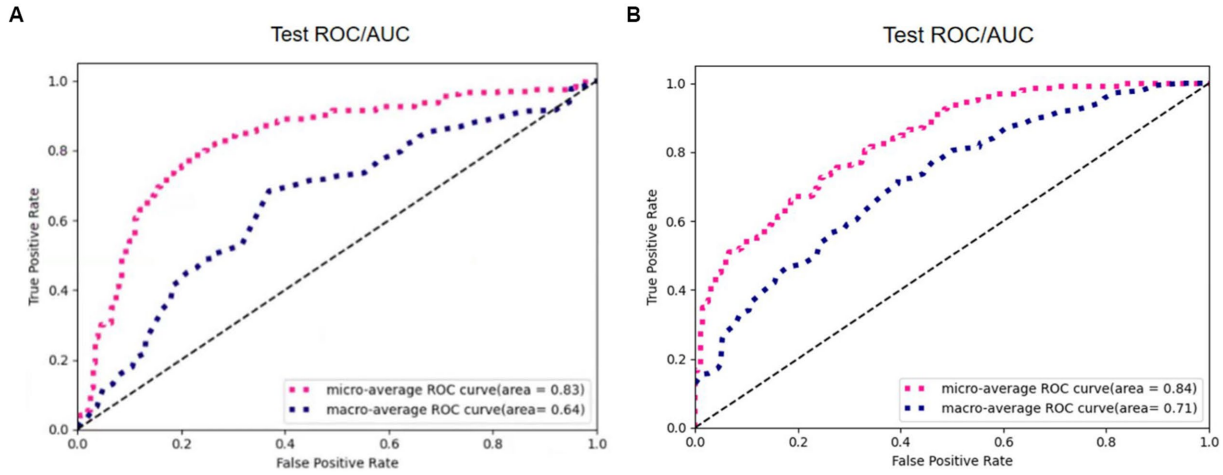


FIGURE 13
The results of 3D Resne50+Radiomics with different threshold algorithms. **(A)** Manual thresholding algorithms. **(B)** Otsu thresholding algorithms.

thresholding algorithm filters the tissue information around the small pulmonary solid nodules and reduces the interference of useless information. In addition, 3D visual and radiomic features are integrated to prevent missing features, and extensive experiments demonstrate the feasibility and interoperability of the two methods.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MM proposed the main idea and did all the experiments. ZY contributed to part of the idea and experiments' analysis. YZ improved the idea and provided datasets and experimental evaluation. All authors contributed to the article and approved the submitted version.

Funding

Ph. D. Startup Fund of Hubei University of Technology.

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

Al-Shabi, M., Lan, B. L., Chan, W. Y., Ng, K. H., and Tan, M. (2019). Lung nodule classification using deep local-global networks. *Int. J. Comput. Assist. Radiol. Surg.* 14, 1815–1819. doi: 10.1007/s11548-019-01981-7

Huang, H., Wu, R., Li, Y., and Chao, P. (2022). Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic CT. *IEEE J. Biomed. Health Inform.* 26, 3860–3871. doi: 10.1109/JBHI.2022.3171851

Kang, G., Liu, K., Hou, B., and Zhang, N. (2017). 3D multi-view convolutional neural networks for lung nodule classification. *PLoS One* 12:e0188290. doi: 10.1371/journal.pone.0188290

Liu, M., Zhang, F., Sun, X., Yu, Y., and Wang, Y. (2021). "CA-Net: leveraging contextual features for lung cancer prediction," In Proceedings of the Medical Image Computing and Computer Assisted Intervention Society, (Strasbourg, French: Springer), 23–32

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., et al. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE Trans. Med. Imaging* 35, 1160–1169. doi: 10.1109/TMI.2016.2536809

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., et al. (2017). Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recogn.* 61, 663–673. doi: 10.1016/j.patcog.2016.05.029

Shi, F., Chen, B., Cao, Q., Wei, Y., Zhou, Q., Zhang, R., et al. (2021). Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest CT images. *IEEE Trans. Med. Imaging* 41, 771–781. doi: 10.1109/TMI.2021.3123572

Sori, W. J., Feng, J., and Liu, S. (2019). Multi-path convolutional neural network for lung cancer detection. *Multidim. Syst. Sign. Process.* 30, 1749–1768. doi: 10.1007/s11045-018-0626-9

Tsai, C.H., and Peng, Y.S. (2022). "Multi-task lung nodule detection in chest radiographs with a dual head network," In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2022, (Singapore: Springer), 707–717

Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., et al. (2018). Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. *IEEE Trans. Med. Imaging* 38, 991–1004. doi: 10.1109/TMI.2018.2876510

Zhang, G., Luo, Y., Zhu, D., Xu, Y., Sun, Y., and Lu, J. (2018). "Spatial pyramid dilated network for pulmonary nodule malignancy classification," In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), (Beijing, China: IEEE), 3911–3916

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 July 2023
DOI 10.3389/fnins.2023.1137557

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Saneera Hemantha Kulathilake,
Rajarata University of Sri Lanka, Sri Lanka
Ling Xing,
Southwest University of Science and Technology, Luoyang, China

*CORRESPONDENCE
Zhiguo Zhou
✉ zhiguozhou@bit.edu.cn
Junwei Duan
✉ jwduan@jnu.edu.cn

RECEIVED 04 January 2023
ACCEPTED 26 June 2023
PUBLISHED 11 July 2023

CITATION
Ma P, Wang J, Zhou Z, Chen CLP, the Alzheimer's Disease Neuroimaging Initiative and Duan J (2023) Development and validation of a deep-broad ensemble model for early detection of Alzheimer's disease.
Front. Neurosci. 17:1137557.
doi: 10.3389/fnins.2023.1137557

Development and validation of a deep-broad ensemble model for early detection of Alzheimer's disease

Peixian Ma[1], Jing Wang[2], Zhiguo Zhou[3]*, C. L. Philip Chen[4], the Alzheimer's Disease Neuroimaging Initiative[5] and Junwei Duan[1,6]*

[1]College of Information Science and Technology, Jinan University, Guangzhou, China, [2]College of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China, [3]School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China, [4]School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, [5]Steering Committee of Alzheimer's Disease Neuroimaging Initiative, Bethesda, MD, United States, [6]Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Guangzhou, China

Introduction: Alzheimer's disease (AD) is a chronic neurodegenerative disease of the brain that has attracted wide attention in the world. The diagnosis of Alzheimer's disease is faced with the difficulties of insufficient manpower and great difficulty. With the intervention of artificial intelligence, deep learning methods are widely used to assist clinicians in the early recognition of Alzheimer's disease. And a series of methods based on data input with different dimensions have been proposed. However, traditional deep learning models rely on expensive hardware resources and consume a lot of training time, and may fall into the dilemma of local optima.

Methods: In recent years, broad learning system (BLS) has provided researchers with new research ideas. Based on the three-dimensional residual convolution module and BLS, a novel broad-deep ensemble model based on BLS is proposed for the early detection of Alzheimer's disease. The Alzheimer's Disease Neuroimaging Initiative (ADNI) MRI image dataset is used to train the model and then we compare the performance of proposed model with previous work and clinicians' diagnosis.

Results: The result of experiments demonstrate that the broad-deep ensemble model is superior to previously proposed related works, including 3D-ResNet and VoxCNN, in accuracy, sensitivity, specificity and F1.

Discussion: The proposed broad-deep ensemble model is effective for early detection of Alzheimer's disease. In addition, the proposed model does not need the pre-training process of its depth module, which greatly reduces the training time and hardware dependence.

KEYWORDS

deep-broad ensemble model, Alzheimer's disease, early detection, MRI, validation, efficiency

1. Introduction

Alzheimer's disease (AD) is a chronic neurodegenerative disease of the brain that develops insidiously. People diagnosed with Alzheimer's will suffer from the disease for remaining lifespan (Todd et al., 2013; Weiner et al., 2013; Fink et al., 2020). The main symptoms of AD includes memory impairment, executive dysfunction, aphasia, impairment of visuospatial skills and so on, and the etiology remains unknown (Mayeux and Sano, 1999; Mimura and Yano, 2006). Thus, Millions of people around the world suffer from Alzheimer's

disease. The long-term treatment of these patients consumes huge medical resources and costs (Cummings and Cole, 2002; Scheltens et al., 2016). The diagnosis of AD can be divided into three main types: AD (Alzheimer's disease), MCI (Mild Cognitive Impairment), NC (Normal Control).

Radiographic images are important in medical diagnosis of Alzheimer's disease. These include positron emission tomography (PET), magnetic resonance imaging (MRI), computed tomography (CT) and so on (Prince and Links, 2006; Doi, 2007; Johnson et al., 2012). Due to low cost and high efficiency, MRI imaging play an important role in diagnosing AD related pathological brain changes and researching (Jack et al., 1999). The understanding of the pathological information provided by these radiographic images depends on the knowledge and experience of the front-line clinicians. As the number of professional medical staff is far less than the actual patient treatment needs, they can not timely diagnose some early hidden symptoms of Alzheimer's disease. At the same time, the imbalance of medical resources also leads to the inability of patients in rural areas to obtain effective early diagnosis locally for follow-up treatment.

Currently, with the proposal of VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), ResNest (Zhang et al., 2022), ResNext (Xie et al., 2017) and a series of deep neural networks, many medical and artificial intelligence researchers used these model to conduct corresponding training on radiographic images. The popularity of high-performance hardware maked it possible to deploy these frameworks in some large hospitals and enable medical departments to actually use these methods to assist physicians in clinical diagnosis and reduce patient care costs. Due to the complex 3-Dimensional (3D) spatial feature of radiographic images, which is extremely different from the traditional 2-Dimensional (2D) images, a variety of model were proposed based on different inputs. Compared to 2D-input models, 3D-input models get more structure information from data obtained by continuous scanning, thus they can extract more complex three-dimensional spital feature. Consequently, in most application scenarios, 3D-input models performs better than the former in recognition tasks. These methods can be divided into 2D deep learning method and 3D deep learning method. The 2D method mainly divides the original medical image into multiple slices on a specific axis and then inputs them into the classical convolutional neural network for training.However, the 2D method cannot learn the correlation feature between these slices, so the model performance is limited. The 3D method directly input the original image into the 3D improved convolutional neural network for training, in order to learn more comprehensive feature information and make up for the above defects.

However, deep models contained a large number of hyperparameters, which required huge hardware resources. The gradient descent method is also prone to fall into the optimal solution, leading to the failure of weight. Researchers need to find a quick and effective way to solve this problem. In recent years, on the basis of Random Vector Functional-Link Neural Network (RVFLNN) (Pao et al., 1994) and Single-layer Linear Feedforward Network (SLFN) (Sanger, 1989), Chen et al. proposed the Broad Learning System (BLS) (Chen and Liu, 2017a,b) and proved its approximation. BLS showed good accuracy and excellent calculation speed in various classification tasks.

Therefore, on the basis of BLS, we try to combine it with deep learning to establish a depth-broad ensemble model. The 3D deep convolution module will enable the model to have the capacity to initially extract features of 3D inputs, while the broad learning module, as a key part of feature fitting, greatly reduces the resource consumption of the model and can maintain a good performance. While Alzheimer's image recognition is a emblematic 3D image processing task, it has had a profound influence on medicine and computer science. There have been a lot of research on the application of depth model in this aspect. Applying our proposed depth-broad ensemble model to the early detetion of Alzheimer's disease will help drive technological innovation, reduce the cost of future applications, and better facilitate the adoption of machine learning technologies in this field.

In this study, we proposed an improved deep-broad ensemble model for the detection of AD. This model combined the 3D extraction capability with the fast operation speed and low dependence on hardware. It firstly extracted spatial features of different levels of images, and then fused multi-level features based on a novel BLS to get better classification results. We applied this model to the task of MRI image recognition in Alzheimer's disease and compared it with some previous work and the work of radiology readers. Experimental results demonstrate that the proposed model has excellent accuracy and computational efficiency.

The main contributions of our study for the early detection of AD can be reported as follows:

1. We constructed a novel deep-broad ensemble model based on 3D residual convolution module and Broad Learning System.
2. The proposed model outperforms previous single deep models, and has higher training efficiency, less dependence in hardwares.
3. There is no need to pre-train the deep modules of the proposed, which greatly reduces the training time.

2. Related works

Early detection of Alzheimer's disease is a chronic and significant research topic in computer science area. At present, a large number of computer-aided early detection methods for Alzheimer's disease have been developed. According to the dimension division of input data, related works can be divided into two-dimensional input-based research methods and three-dimensional input-based research methods. Two-dimensional input methods consist of traditional machine learning model and two-dimensional deep learning models. The three-dimensional input method basically takes the three-dimensional deep learning model as the main backbone. Due to the abundant spatial pathological information in medical examination images, three-dimensional methods generally have more advantages in the detection effect. In addition, some researchers have studied the characteristics of small sample size of medical test images, or introduced different types of data to establish a multi-modal fusion algorithm.

Rieke et al. trained on a 3D CNN model and applied four gradient-based and occlusion-based approaches to visualization, promoting clinical impact and trust in computer-based decision support systems (Rieke et al., 2018). But 3DCNN contains the risk of network degradation and gradient disappearance/gradient explosion after increasing the number of middle layers. Rieke et al. trained on a 3D CNN model and applied four gradient-based and occlusion-based approaches to visualization, promoting clinical impact and trust in computer-based decision support systems. But 3DCNN contains the risk of network degradation and gradient disappearance/gradient explosion after increasing the number of middle layers. Based on convolutional autoencoder (CAE), Kanghan et al. proposed supervised and unsupervised classification methods for the diagnosis of Alzheimer's disease. The combination of convolutional layer and pooling layer of CAE is relatively fixed, which limits to construct more complex network structure. Guan et al. constructed a preliminary standardized model framework based on ResNet, VGG, DenseNet and other networks, and comprehensively tested and compared these models using standard MRI image data sets of Alzheimer's disease. They found that these simple architectures performed similarly on the task, and the pre-training process of these methods has less impact on accuracy. Korolev et al. proposed VoxCNN based on ResNet to classify MRI images. This model can achieve better performance using a small training dataset, and be applied to 3D MRI images without the need of intermediate handcrafted feature extraction. However, VoxCNN includes the module of 3D-Resnet, which needs to consume more training time and more computing resources in training.

In the above reports, these methods have achieved excellent performance in their selected datasets. However, these studies lacked comparability and robustness among themselves. Most studies needs lots of GPU resources to train the model, which makes it difficult to apply the research results widely.

3. Methods

3.1. Approvement statement of institutional review board

This study is approved by institutional board with written informed consent waived. All experiments including any relevant details are approved by institutional and/or licensing committee. All experiments on humans and/or the use of human tissue samples were performed in accordance with relevant guidelines and regulations. All experimental protocols were approved by the Steering Committee of Alzheimer's Disease Neuroimaging Initiative and Academic Committee of Jinan University. Informed consent was obtained from all subjects and/or their legal guardian(s).

3.2. Data acquisition

All MRI image data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/) (Petersen et al., 2010). Founded in 2003, ADNI is a public-private partnership led by Principal Investigator Michael W.Weiner, MD. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biomarkers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The latest information about the ANDI database can be found at (http://adni.loni.usc.edu/). The ADNI database consists of four sub-databases, including ADNI-1, ADNI-Go, ADNI-2, and ADNI-3, which are interdependent. The diagnostic labels of these medical image data are given by doctors after a series of tests.

ADNI provides several standardized datasets for researchers to study. In our research, ADNI1 Complete 2Yr 3T standardized dataset is chosen to train our model, including scans of patients taken at 6, 12, 18, and 24 months after diagnosis. The dataset contains 434 subjects, including 77 of AD, 206 of MCI, and 151 of NC. The demographic information of the dataset is shown in Table 1. We searched the ADNI1 Complete 2Yr 3T standardized dataset in the ADNI database, packaged it and downloaded. All image data in this study were stored in NifTI format. Figure 1 shows the slides samples of this dataset.

3.3. Data preprocessing

Since the data came from many different patient samples, the size of different data and the location of key parts in the image may vary to some extent, while the neural network model required that the size of each input be consistent. At the same time, due to its intensity and other attributes, MRI image is not suitable to be directly used as the input image of the model, so it needs to be transformed to some extent. In conclusion, we have to take a series of pre-processing measures for the data, so that it can be converted into appropriate input data, and is conducive to improving the performance of the model.

In this research, The primitive image size of our ADNI dataset was $256\times256\times160$. Firstly, Since the pixel size of different medical scanned images is not the same, all pixels need to be resampled at a fixed homogeneous resolution. We resampled all MRI images to 1.5-mm isotropic voxels. Then, we scaled the intensity of the images to the range (0, 1). Since the region of interest (ROI), namely the patient's brain, was basically concentrated in the center of the image, we cropped the image based on the central region and removed some peripheral background areas. The final output after processing were $224\times224\times128$-pixel grid resulting in $336\times336\times192$-mm^2 volume. The above data preprocessing operations were based on Python 3.8 environment, using package Monai (https://monai.io/) and package Numpy (https://numpy.org/) for processing.

3.4. Model details and training

3.4.1. Broad learning system

Although traditional deep neural networks show good accuracy in traditional recognition and classification tasks, they also expose problems such as large number of hyperparameters and long time consuming. At the same time, with the publication of more types of datasets, researchers need to seek a new method

TABLE 1 The structure of residual Conv module used in feature mapping layer and enhancement layer.

Layer name	Number of Bottle Neck block	Number of kernel	Kernel size	Input Size	Output size
3D Conv Layer	0	64	7×7×7	224×224×128	224×112×64
3D AvgPool Layer	0	64	3×3×3	224×112×64	112×56×32
3D Residual Module 1	3	256	1×1×1 3×3×3 1×1×1	112×56×32	112×56×32
Global AvgPool	0	256	0	112×56×32	256
3D Residual Module 2	4	512	1×1×1 3×3×3 1×1×1	112×56×32	56×28×16
Global AvgPool	0	512	0	56×28×16	512

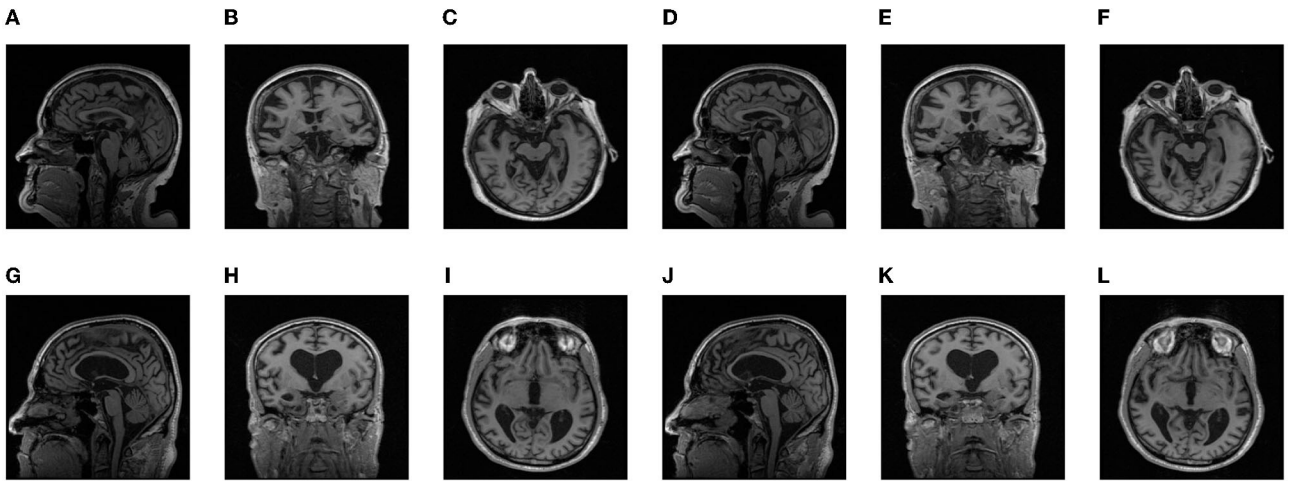


FIGURE 1
Partial slice samples of some ADNI datasets: **(A–F)** AD, **(G–L)** NC.

with simple structure and fast operation to deal with different requirements and tasks. In studies over the past few years, classical single-layer network structures such as Extreme Learning Machine (ELM) (Huang et al., 2006) and Random Vector Functional Link Neural Network (RVFLNN) (Pao et al., 1994) have been proposed successively.

On the basis of RVFLNN (Pao et al., 1994), Chen et al. proposed Broad Learning System (BLS) (Chen and Liu, 2017a,b). In the structure of BLS, the basic linear feature of the input was extracted by the feature mapping layer. The further feature of the former layer was extracted by the enhancement layer which contained a non-linear function. Then, the output of these layers were concat together and transferred to the output layer for classifying. Since there is only two layers of structure, BLS does not need to calculate a large number of weight parameters for multiple middle layers, which saving a lot of calculation resources and reducing the training time of the model. Previous experimental results demonstrate that BLS can still achieve excellent performance in the basic test of image recognition, which proves that BLS has good potential in the field of computer vision (Chen and Liu, 2017a,b).

For a given input sample $\boldsymbol{X} \in \boldsymbol{R}^{n \times m}$, where $\boldsymbol{n}$ represents the number of samples, $\boldsymbol{m}$ represents the feature dimension of the sample. The feature mapping layer is composed of the combination of feature nodes. The feature nodes and feature mapping layer of broad learning system can be expressed as following:

$$\boldsymbol{d_i} = \boldsymbol{\varphi_i}(\boldsymbol{X}\boldsymbol{W_{e_n}} + \boldsymbol{\beta_{e_n}}) \quad (1)$$

$$\boldsymbol{D^n} = [\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots\ldots \boldsymbol{d_n}] \quad (2)$$

where $\boldsymbol{\varphi_i}$ is the selectable linear or non-linear activation function, $\boldsymbol{W_{e_n}}$ is the random weight, and $\boldsymbol{\beta_{e_n}}$ is the random bias. $\boldsymbol{W_{e_n}}$ and $\boldsymbol{\beta_{e_n}}$ are usually optimized by sparse auto-coding algorithm. The enhancement node and enhancement layer of BLS can be denoted as:

$$\boldsymbol{e_j} = \boldsymbol{\delta_j}(\boldsymbol{D^n}\boldsymbol{W_{h_m}} + \boldsymbol{\beta_{h_m}}) \quad (3)$$

$$\boldsymbol{E^m} = [\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots\ldots \boldsymbol{E_m}] \quad (4)$$

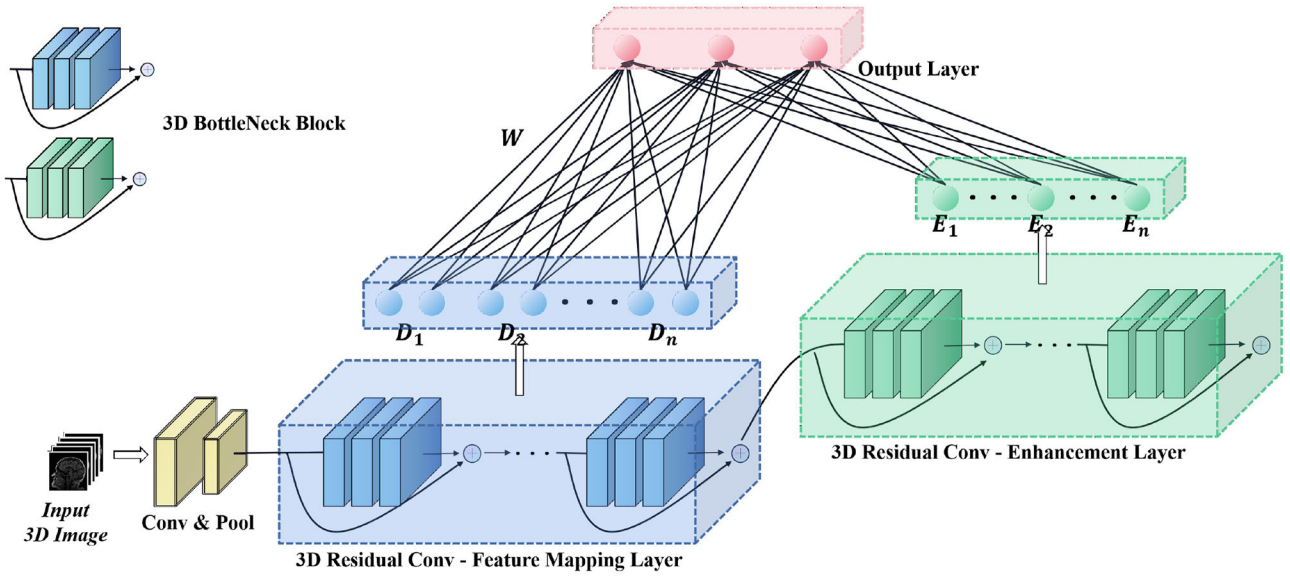


FIGURE 2
Structure of proposed deep-broad ensemble model.

TABLE 2 Demographics of ADNI dataset.

Diagnosis	Number of patients	Number of images	Gender Male/Female	Average age All [Male/Female]
AD	18	77	32/45	75.32 [75.71/75.04]
NC	33	151	58/93	76.62 [77.46/76.08]
MCI	35	206	132/74	74.62 [77.45/69.56]
Total	86	434	222/212	75.44 [77.20/73.58]

where δ_j is the non-linear activation function, W_{h_m} is the random weight, and β_{h_m} is the random bias. Then, the feature mapping layer and the enhancement layer are concated and transferred to the output layer. Since W_{e_n}, β_{e_n}, W_{h_m}, and β_{h_m} remain unchanged in the training process, the objective function of BLS is:

$$_W(||(Y - Y')||_2^2 + \frac{\lambda}{2}||W||_2^2) \quad (5)$$

where W is the weight of the output layer of the BLS, Y is the label of X, $||(Y - Y')||_2^2$ is used to control the minimization of training error, $\frac{\lambda}{2}||W||_2^2$ is used to prevent model overfitting, and λ is the regularization coefficient. Then, W can be obtained by seeking the pseudo-inverse of ridge regression:

$$W = G^+ Y \quad (6)$$

$$G^+ = \lim_{\lambda \to 0} (\lambda I + G^T G)^{-1} G^T \quad (7)$$

where I is the identity matrix. Through the above steps, we constructed a complete Broad Learning System.

3.4.2. Deep-broad ensemble model

Three-dimension radiographic images contain complex pathological spatial information. However, the original broad learning system can only receive two-dimensional features as input, and the ability to extract complex image features is weak. Integrating the deep convolution module can effectively improve the feature extraction ability of the broad learning system, which can better the performance for its classification and recognition (Chen et al., 2018).

In this paper, we proposed a deep-broad ensemble model for early recognition of Alzheimer's disease based on the above ideas, which aims to maintain a considerable performance of classification, reduce the dependence of hardware and improve the efficiency of the model.

As shown in Figure 2, we used a convolution-pooling layer to initially extract the features of the original input. The size of the convolution kernel used was 7×7×7, and the size of the pooling module was 3×3×3.The backbone of the model is composed of a 3D residual convolution—feature mapping layer and a 3D residual convolution—enhancement layer. The 3D residual convolution—feature mapping layer can be divided into residual convolution module and feature mapping module. The former is composed of several 3D bottleneck convolution modules (He et al., 2016), which are used to extract shallow features of the input, and transform these features into feature vectors with a size of 256 through the global pooling, and then input into the feature mapping module for further processing; The 3D residual convolution enhancement layer also includes several 3D bottleneck convolution modules as the

former to extract deeper features. The feature vectors are then input to the enhancement module through the global pooling, which is different from the input of the enhancement layer in the original BLS. Detailed parameters of the deep module used in the whole backbone model are shown in Table 1. Finally, feature mapping module and enhancement module are mapped to the output layer to produce classification results.

For a Given the original MRI image input $\boldsymbol{X}$, the output feature vector after the first convolution-pooling layer is:

$$\boldsymbol{X_{base}} = \boldsymbol{\lambda_{conv-pool}(X)} \tag{8}$$

The feature vector of the output after the residual convolution module of the 3D residual convolution-feature mapping layer and the 3D residual convolution-enhancement layer can be denoted as:

$$\boldsymbol{X_d} = \boldsymbol{\lambda_d(X_{base})} \tag{9}$$

$$\boldsymbol{X_e} = \boldsymbol{\lambda_e(\lambda_d(X_{base}))} \tag{10}$$

where $\lambda_d()$ indicates the residual convolution module of 3D residual convolution-feature mapping layer, and $\lambda_e()$ indicates the residual convolution module of 3D residual convolution-enhancement layer. According to formula (1)–(4), feature nodes, feature mapping modules, enhancement nodes, and enhancement mapping modules of Broad learning can be denoted as:

$$\boldsymbol{d_i} = [\boldsymbol{\varphi_i(X_d W_{e_n} + \beta_{e_n})}] \tag{11}$$

$$\boldsymbol{D^n} = [\boldsymbol{d_1, d_2, ...d_n}] \tag{12}$$

$$\boldsymbol{e_j} = [\boldsymbol{\delta_j(X_e W_{h_m} + \beta_{h_m})}] \tag{13}$$

$$\boldsymbol{E^m} = [\boldsymbol{E_1, E_2, ...E_m}] \tag{14}$$

According to formula (6)–(7), the final classification output $\boldsymbol{Y}$ can be obtained. Thus, we constructed a 3D Convolution Broad Learning System based on 3D medical image input.

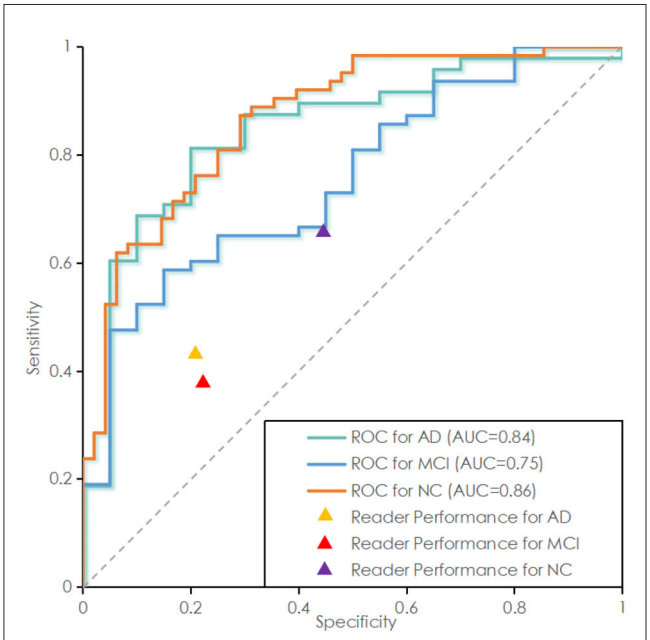


FIGURE 3
Receiver operating characteristic (ROC) curve of the proposed deep-broad ensemble model trained on 70% of 3YR 2T ADNI dataset and tested on the remaining 30% of ADNI dataset and independent test set. ROC curve labeled Alzheimer Disease (AD) represents the essential performance for distinguishing AD vs. all other cases. ROC curves for Mild Cognitive Impairment (MCI) and Normal Control (NC) are also reported for technical completeness.

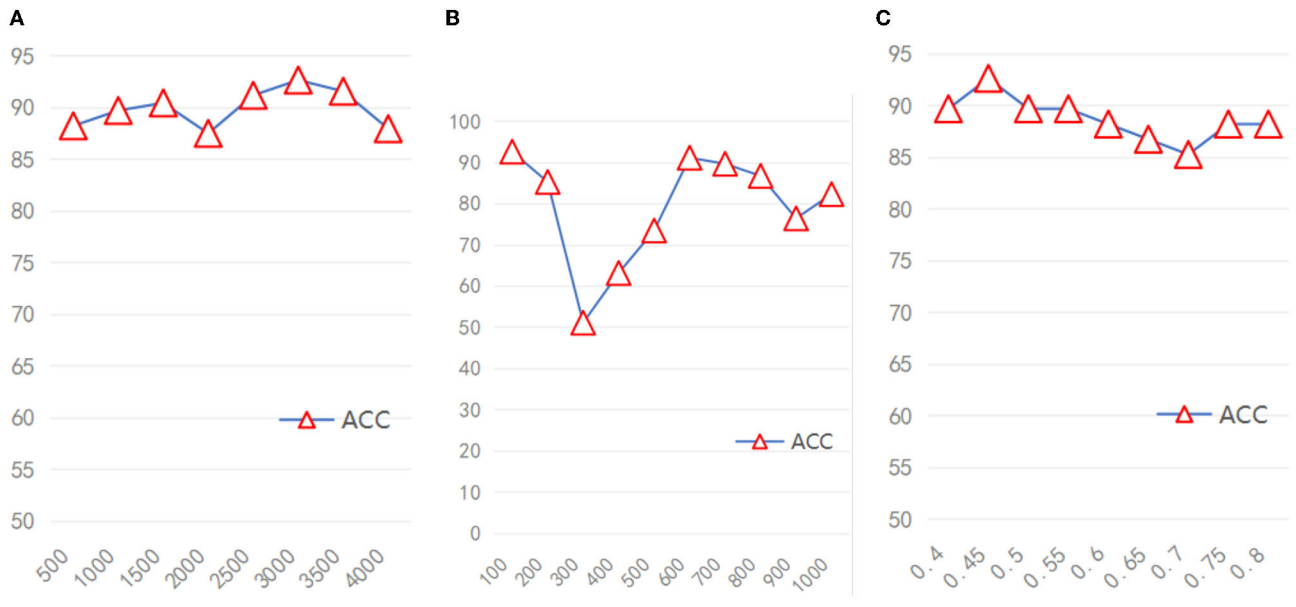


FIGURE 4
The accuracy of the model under different hyperparameters **(A)** Feature nodes, **(B)** Enhancement nodes, **(C)** Sparsity coefficient.

3.4.3. Model training

After data preprocessing, the original image was converted into a 224×224×128 matrix as the input of the model. The server we used for model training had a AMD EPYC 7543 32-core Processor, 90 GB of RAM, Nvidia GeForce RTX 3090 GPU with CUDA 11.1.

We set up two groups of experiments whose convolution module was non-pre-trained and pre-trained respectively for comparison. The dataset used for pre-training was ADNI training dataset (347 images). In the pre-training experiment, we set cross entropy as the loss function, with the learning rate of 0.0001, the optimizer of momentum-SGD, and the batch size of 4.

For the broad module in the model, the range of feature nodes was [500–4,000], the range of enhancement nodes is [100–1,000], and the range of sparsity coefficient is [0.4–0.7]. We used Pytorch 1.8 and Numpy to construct the program of the model presented in this article, and all programs and experiments were run in Python 3.8.

TABLE 3 Comparison of the proposed model and radiology readers.

Method	Accuracy (%)	Sensitivity (%)	Precision (%)
AD vs. NC			
Deep-broad ensemble model (Pre-trained)	90.57	91	91
Deep-broad ensemble model (Not Pre-trained)	92.65	91	91
AD vs. MCI			
Deep-broad ensemble model (Pre-trained)	93.58	92	91
Deep-broad ensemble model (Not Pre-trained)	91.57	92	92
MCI vs. NC			
Deep-broad ensemble model (Pre-trained)	76.44	74	75
Deep-broad ensemble model (Not Pre-trained)	84.68	85	84

The data are presented as Maximum.

3.5. Model testing and analysis

For the three groups of models after training, we used the ADNI test dataset for testing. The model finally outputs the probability that an image belongs to one of these categories. The category with the highest probability was selected as the classification result. We calculated the final classification accuracy, Precision and Recall based on this result. In addition, We studied the stability of the model by modifying the hyperparameters.

3.6. Clinical interpretation of MRI

To compare the performance of our proposed model with that of an actual radiology reader, a board-certified nuclear medicine physician with several years of experience (HuanHua Wu, nuclear medicine) was invited to perform a discriminative analysis of 87 MRI images from the ADNI test dataset. In order to prevent data leakage, the reader can only obtain MRI image data and the number of the subject, and analyze them based on their professional experience. We will calculate the corresponding indicators based on this result.

4. Results

4.1. Demographics

As shown in Table 2, The dataset used in this study contained 434 MRI images from 86 patients, which contained three types of Alzheimer's symptoms: AD, patients with Alzheimer's disease; MCI, mild cognitive impairment; NC, normal person. Seventy-seven images were obtained from AD, 151 from NC, and 206 from MCI. Partial slice images of AD and NC cases in the dataset are shown in Figure 4. The average age of all patients was 75.44 years old (range from 55 to 90 years), including 73.58 years old for female (range from 55 to 90 years) and 77.20 for male (range from 57 to 89 years). The average age of AD groups was 75.32 years (range from 57 to 90 years), with the average age of 75.04 years for female (range from 64 to 90 years) and 75.71 years for male (range from 57 to 87 years). The average age of MCI groups was 74.62 years (range from 55 to 89 years), with the average age of 69.56 years for female

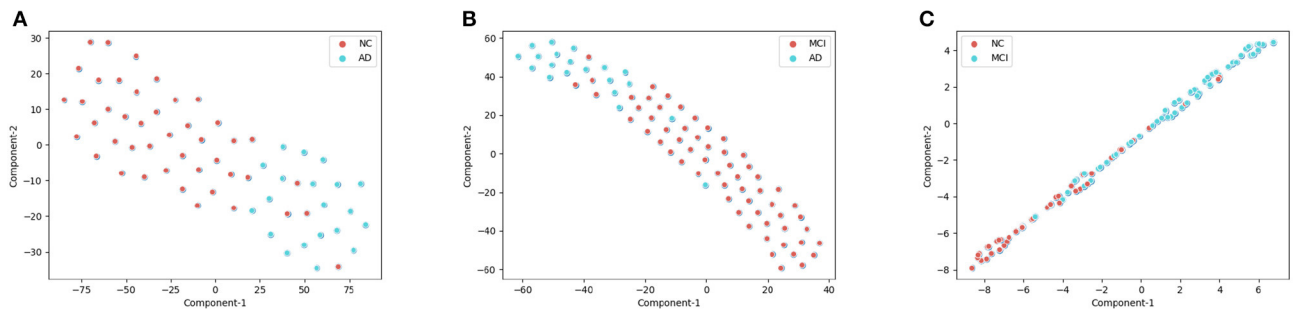


FIGURE 5
The visualization of training set after dimension reduction with t-distributed stochastic neighbor embedding (t-SNE). Each point represents the final features of the proposed deep-broad ensemble model. **(A)** t-SNE for the AD/CN, **(B)** t-SNE for the AD/MCI, **(C)** t-SNE for the NC/MCI.

TABLE 4 Comparison of the proposed model and previous related work.

Method	Accuracy (%)	Sensitivity (%)	F1-score(%)	Time(s)
AD vs. NC				
Deep-broad ensemble model	90.97 ± 1.02	91	91	1.3731
ResNet-18 (He et al., 2016)	73.53 ± 0.97	74	67	†1920
DenseNet-121 (Huang et al., 2017)	69.12 ± 2.93	69	69	†2480
VCNet (Rieke et al., 2018)	70.56 ± 2.91	71	58	†1440
CAE (Oh et al., 2019)	85.24 ± 3.97	88	nan	nan
ICAE (Oh et al., 2019)	86.60 ± 3.66	88	nan	nan
Guan's work (Guan et al., 2019)	69.12 ± 0.57	69	64	†2120
VoxCNN (Korolev et al., 2017)	72.06 ± 1.43	72	64	†2800
AD vs. MCI				
Deep-broad ensemble model	91.16 ± 3.86	94	94	1.4745
ResNet-18 (He et al., 2016)	77.11 ± 1.20	77	68	†1920
DenseNet-121 (Huang et al., 2017)	70.59 ± 1.47	71	58	†2480
VCNet (Rieke et al., 2018)	74.69 ± 1.24	75	65	†1440
CAE (Oh et al., 2019)	74.68 ± 6.04	75	nan	nan
ICAE (Oh et al., 2019)	75.06 ± 3.86	77	nan	nan
Guan's work (Guan et al., 2019)	71.08 ± 2.41	71	67	†2120
VoxCNN (Korolev et al., 2017)	75.90 ± 0.10	76	66	†2800
MCI vs. NC				
deep-broad ensemble model	83.39 ± 1.31	85	84	1.6825
ResNet-18 (He et al., 2016)	57.65 ± 0.90	58	55	†1920
DenseNet-121 (Huang et al., 2017)	55.86 ± 2.81	56	45	†2480
VCNet (Rieke et al., 2018)	58.56 ± 0.91	59	53	†1440
CAE (Oh et al., 2019)	62.83 ± 5.17	66	nan	nan
ICAE (Oh et al., 2019)	63.34 ± 4.16	69	nan	nan
Guan's work (Guan et al., 2019)	56.76 ± 3.72	58	57	†2120
VoxCNN (Korolev et al., 2017)	59.46 ± 0.90	59	49	†2800

Unless otherwise stated, the data are presented as Mean ± Std.
†The data here is represented as the average of the three tasks.

(range from 55 to 82 years) and 77.45 years for male (range from 63 to 89 years). The average age of NC groups was 76.62 years (range from 70 to 88 years), with the average age of 76.08 years for female (range from 71 to 82 years) and 77.46 years for male (range from 70 to 88 years).

4.2. Result of training

The preprocessed dataset was divided into training set and test set in a ratio of 0.7:0.3. We trained models for AD/CN, AD/MCI, and MCI/NC tasks respectively. Accuracy (ACC), sensitivity (SEN), and F1-score were used to evaluate the performance of them, and the training time of each model was recorded.

As shown in Table 4, in the tests of AD/CN, AD/MCI, and MCI/NC, the average of accuracy for prediction were 90.97, 91.16, and 83.39%. SEN is 91, 94, 85%. F1-score is 91, 94, 84%. The above results indicated that the proposed model has good discrimination ability in AD/NC and AD/MCI tasks, but weak discrimination ability in MCI/NC tasks. The ROC curves of the proposed deep-broad ensemble model method trained on 70% of ADNI dataset were shown in Figure 3.

Because the broad module of the model required different hyperparameters, We also verified the stability of the model based on different hyperparameters. Due to the huge range of hyperparameters (range of feature mapping nodes: 500–4,000, range of enhancement node: 100–1,000, range of sparsity coefficient: 0.4–0.7), we took several hyperparameters values as representative. As shown in Figures 4A–C, when the number of feature mapping nodes and the sparse coefficient increased, the model maintained good stability. When the number of enhancement nodes increases, the stability of the model is generally acceptable, excluding some unstable intervals.

TABLE 5 Comparison of the proposed model and radiology readers.

Method	Accuracy (%)	Sensitivity (%)	Specificity	F1-score(%)
AD vs. NC				
Deep-broad ensemble model	92.65	91	89	91
Radiology readers	68.57	38	22	35
AD vs. MCI				
deep-broad ensemble model	93.58	94	92	94
Radiology readers	75.00	43	21	29
MCI vs. NC				
deep-broad ensemble model	84.68	85	81	84
Radiology readers	61.76	66	45	68

The data are presented as Maximum.

In addition, We explored the influence of the pre-training for the deep module on the final classification performance of the model. As shown in Table 3, the non-pre-trained model performed better than the pre-trained model on the AD/NC and NC/MCI tasks (maximum accuracy of 92.65/90.57 and 84.68/76.44), while the two performed similarly on the AD/MCI tasks (maximum accuracy of 92.76/91.57).

Since the BLS itself had considerable ability of feature fitting, the deep module of the proposed model was mainly used to further extract complex space features of medical image, enhancing the feature extraction ability of BLS. Therefore, the pre-training of deep module was not decisive. The broad module is decisive in the fitting of image features. If the cost of pre-training process was removed, the training time of the model proposed can be further reduced, lowing the dependency of hardwares and improving the efficiency of the model.

4.3. Model interpretation: t-SNE plot

As shown in the Figure 5, We clustered the final features of the models in the three experiments respectively after dimension reduction by t-SNE. In AD/NC and AD/MCI classification experiments, the corresponding categories were almost pure, with only a small amount of mixing. In the NC/MCI experiment, the mixture of the two categories was more common. Therefore, we concluded that the proposed model is highly sensitive to AD categories, because most of the sample points were in the clustering of AD; we achieved high accuracy in both experiments.

4.4. Comparison to previous works

We compared the proposed model with some previous works, which had developed several deep models in this task. Due to the lack of relevant hyperparameter reference, we set the number of training epochs to 40 for each work. As shown in Table 4, in the three tasks of AD/CN, AD/MCI and MCI/NC, the accuracy of the proposed model outperformed these works. The training time of the proposed model was much shorter than that of these deep models, because there was no need to update the weight parameters of the deep module of the proposed model. Therefore, compared with the previous works, the proposed model had a considerable optimization effect, less dependence on computer hardware, and was easier to deploy in the actual diagnosis process.

4.5. Comparison to clinical Interpretations

As shown in Table 5, in the above tasks, the accuracy of radiology readers were 68.57, 75.00, 61.76%. Sensitivity were 38, 43, 66%. F1-score were 35, 29, 68%. Compared to radiology reader's work, the proposed model had better performance in the detection of ADNI datasets, which has statistical significance.

5. Discussion

The diagnosis and treatment of Alzheimer's disease is becoming an important medical issue for decades to come. Millions of Patients with Although Alzheimer's disease provides a rich data base for the improvement of diagnostic theories, it brings great work pressure and challenges to front-line doctors.

At present, computer science researchers had developed many detection models for radiographic images of the Alzheimer's disease. However, these current models almost consisted of single deep networks. This would lead to the problem that the models were highly dependent on hardwares, which was difficult to be popularized in non-urban areas where relevant hardware was lacking. To solve the above problems, we constructed a deep-broad ensemble model for radiographic images based on the novel BLS which has higher efficiency. Then, we trained and tested the model using the MRI dataset obtained from ADNI database, and calculated the corresponding accuracy, sensitivity and F1-score according to the results. We also compared the model with some previous work and results from radiology reader. The results demonstrates that compared with the previous work and the reader, the proposed model has better performance and greatly reduces the training time. Meanwhile, we studied the effect of the deep convolution module and the improved BLS module on the model. The results demonstrates that BLS was still the core

of the model. The function of the deep module was to enhance the feature extraction capability of the BLS module, thus requiring no pre-training, which would greatly improve the performance of the model proposed in this paper. Our experiment had certain limitations. Firstly, the amount of data used in this study is still relatively small (434 images) due to the limited number of public medical image datasets for Alzheimer's disease currently available for research. Therefore, the robustness of the proposed model has not been verified on larger and more general data, which limits the application of our proposed model in real scenarios.

Second, BLS is a non-deep learning framework. Although its interpretability of it has been proven (Chen et al., 2018), BLS is not as widely used as deep neural network. Broad Learning System itself also has the limitation of relatively low accuracy, and its application in medical imaging and other fields lacks of universal reference. The hyperparameter setting of the proposed model relies on the previous research experience of machine learning researchers and lacks a better adjustment method (Gong et al., 2021).

In general, our experiment and research results demonstrate that our proposed deep-broad ensemble model method significantly reduces the training time while maintaining good detection performance. This makes our model play a referential role in practical medical image diagnosis and reduces the dependence on external hardware. With the opening of more medical image data, the model proposed in this paper can be better applied to first-line clinical diagnosis and provide reliable reference for doctors and medical image readers.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by Academic Committee of Jinan University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

PM: conceptualization, methodology, validation, visualization, software, writing—original draft, review, and editing, and project administration. JW: supervision and project administration. ZZ: supervision and writing—review and editing. CC: conceptualization, resources, and supervision. Alzheimer's Disease Neuroimaging Initiative: data provision. JD: conceptualization, methodology, validation, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018YFC2002500, in part by Guangdong Basic and Applied Basic Research Foundation under Grant No. 2021A1515011999, and in part by Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization under Grant No. 2021B1212040007. In addition, the authors thank the ADNI Committee for providing the MRI data.

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

Chen, C. P., and Liu, Z. (2017a). "Broad learning system: a new learning paradigm and system without going deep," in *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (IEEE), 1271–1276.

Chen, C. P., and Liu, Z. (2017b). Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 10–24. doi: 10.1109/TNNLS.2017.2716952

Chen, C. P., Liu, Z., and Feng, S. (2018). Universal approximation capability of broad learning system and its structural variations. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 1191–1204. doi: 10.1109/TNNLS.2018.2866622

Cummings, J. L., and Cole, G. (2002). Alzheimer disease. *JAMA* 287, 2335–2338. doi: 10.1001/jama.287.18.2335

Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput. Med. Imaging Graph.* 31, 198–211. doi: 10.1016/j.compmedimag.2007.02.002

Fink, H. A., Linskens, E. J., Silverman, P. C., McCarten, J. R., Hemmy, L. S., Ouellette, J. M., et al. (2020). Accuracy of biomarker testing for neuropathologically defined Alzheimer disease in older adults with dementia: a systematic review. *Ann. Internal Med.* 172, 669–677. doi: 10.7326/M19-3888

Gong, X., Zhang, T., Chen, C. P., and Liu, Z. (2021). Research review for broad learning system: algorithms, theory, and applications. *IEEE Trans. Cybernet.* 52, 8922–8950. doi: 10.1109/TCYB.2021.3061094

Guan, Z., Kumar, R., Fung, Y. R., Wu, Y., and Fiterau, M. (2019). A comprehensive study of Alzheimer's disease classification using convolutional neural networks. *arXiv preprint arXiv:1904.07950*.

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.

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*, 4700–4708.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing* 70, 489–501. doi: 10.1016/j.neucom.2005.12.126

Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., et al. (1999). Prediction of ad with MRI-based hippocampal volume in mild cognitive impairment. *Neurology* 52, 1397–1397. doi: 10.1212/WNL.52.7.1397

Johnson, K. A., Fox, N. C., Sperling, R. A., and Klunk, W. E. (2012). Brain imaging in Alzheimer disease. *Cold Spring Harbor Perspect. Med.* 2, a006213. doi: 10.1101/cshperspect.a006213

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). "Residual and plain convolutional neural networks for 3D brain MRI classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (IEEE), 835–838.

Mayeux, R., and Sano, M. (1999). Treatment of Alzheimer's disease. *N. Engl. J. Med.* 341, 1670–1679.

Mimura, M., and Yano, M. (2006). Memory impairment and awareness of memory deficits inearly-stage Alzheimer's disease. *Rev. Neurosci.* 17, 253–266. doi: 10.1515/REVNEURO.2006.17.1-2.253

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019a). Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci. Rep.* 9, 1–16. doi: 10.1038/s41598-019-54548-6

Pao, Y.-H., Park, G.-H., and Sobajic, D. J. (1994). Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing* 6, 163–180. doi: 10.1212/WNL.0b013e3181cb3e25

Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., et al. (2010). Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74, 201–209.

Prince, J. L., and Links, J. M. (2006). *Medical Imaging Signals and Systems*. Upper Saddle River, NJ: Pearson Prentice Hall.

Rieke, J., Eitel, F., Weygandt, M., Haynes, J. -D., and Ritter, K. (2018). "Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease," in *Machine Learning in Clinical Neuroimaging* (Granada).

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Netw.* 2, 459–473.

Scheltens, P., Blennow, K., Breteler, M., De Strooper, B., Frisoni, G., and Salloway, S. (2016). Van der flier WM: Alzheimer's disease. *Lancet* 388, 505–517. doi: 10.1016/S0140-6736(15)01124-1

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

Todd, S., Barr, S., Roberts, M., and Passmore, A. P. (2013). Survival in dementia and predictors of mortality: a review. *Int. J. Geriatr. Psychiatry* 28, 1109–1124. doi: 10.1002/gps.3946

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2013). The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimers Dement.* 9, e111–e194. doi: 10.1016/j.jalz.2013.05.1769

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 1492–1500.

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., et al. (2022). "ResNest: split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 2736–2746.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 17 July 2023
DOI 10.3389/fnins.2023.1159914

OPEN ACCESS

EDITED BY
Adeel Razi,
Monash University, Australia

REVIEWED BY
Xintao Hu,
Northwestern Polytechnical University, China
Zhan Xu,
University of Texas MD Anderson Cancer Center, United States

*CORRESPONDENCE
Shella Keilholz
✉ shella.keilholz@bme.gatech.edu

RECEIVED 06 February 2023
ACCEPTED 05 June 2023
PUBLISHED 17 July 2023

CITATION
Kashyap A, Plis S, Ritter P and Keilholz S (2023) A deep learning approach to estimating initial conditions of Brain Network Models in reference to measured fMRI data. *Front. Neurosci.* 17:1159914. doi: 10.3389/fnins.2023.1159914

A deep learning approach to estimating initial conditions of Brain Network Models in reference to measured fMRI data

Amrit Kashyap[1,2], Sergey Plis[3], Petra Ritter[1,2] and Shella Keilholz[4]*

[1]Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany, [2]Department of Neurology with Experimental Neurology, Brain Simulation Section, Charité – Universitätsmedizin Berlin, corporate member of Freie Universitat Berlin and Humboldt-Universitat zu Berlin, Berlin, Germany, [3]Department of Computer Science, Georgia State University, Atlanta, GA, United States, [4]Wallace H. Coulter Department of Biomedical Engineering, Georgia University of Technology and Emory University, Atlanta, GA, United States

Introduction: Brain Network Models (BNMs) are mathematical models that simulate the activity of the entire brain. These models use neural mass models to represent local activity in different brain regions that interact with each other via a global structural network. Researchers have been interested in using these models to explain measured brain activity, particularly resting state functional magnetic resonance imaging (rs-fMRI). BNMs have shown to produce similar properties as measured data computed over longer periods of time such as average functional connectivity (FC), but it is unclear how well simulated trajectories compare to empirical trajectories on a timepoint-by-timepoint basis. During task fMRI, the relevant processes pertaining to task occur over the time frame of the hemodynamic response function, and thus it is important to understand how BNMs capture these dynamics over these short periods.

Methods: To test the nature of BNMs' short-term trajectories, we used a deep learning technique called Neural ODE to simulate short trajectories from estimated initial conditions based on observed fMRI measurements. To compare to previous methods, we solved for the parameterization of a specific BNM, the Firing Rate Model, using these short-term trajectories as a metric.

Results: Our results show an agreement between parameterization of using previous long-term metrics with the novel short term metrics exists if also considering other factors such as the sensitivity in accuracy with relative to changes in structural connectivity, and the presence of noise.

Discussion: Therefore, we conclude that there is evidence that by using Neural ODE, BNMs can be simulated in a meaningful way when comparing against measured data trajectories, although future studies are necessary to establish how BNM activity relate to behavioral variables or to faster neural processes during this time period.

KEYWORDS

Brain Network Model, fMRI, deep learning, Neural ODEs, initial condition

1. Introduction

Brain Network Models (BNMs) represent whole brain activity as the coordination of many distinct neural populations that are connected via a structural network consisting of long-distance white matter tracts (Sanz-Leon et al., 2015; Breakspear, 2017). Simulations of these network models are being compared to experimental measurements such as functional magnetic

resonance imaging (fMRI). At this spatiotemporal scale in fMRI, the measured activity is thought to be an averaged property of the neural populations and occurs relatively slowly (<1 Hz) compared to the faster neural information processes and the measured fMRI signal is thought to represent the coordination between brain regions over the structural network (Deco et al., 2009). BNMs have been able to reproduce properties observed in fMRI, especially during rest where the brain is not exposed to a structured experimental task or stimulus and the whole brain activity is thought to mostly arise from intrinsic network loops between cortical regions (Honey et al., 2007; Cabral et al., 2011; Sanz-Leon et al., 2015). Thus, BNMs have been used as a generative framework in order to analyze how local neural activity could translate to global coordination and how changes due to neural pathologies translate to observed aberrant dynamics (Ritter et al., 2013; Sanz-Leon et al., 2015; Saenger et al., 2017; Schirner et al., 2018).

Current research does not utilize a single type of population model to construct BNMs. Instead, depending on the application and the underlying assumptions, different neural mass models are selected to represent whole brain activity (Sanz-Leon et al., 2015). For replicating rs-fMRI, several models have been shown to reproduce time-averaged properties such as functional connectivity (FC), computed via cross correlation of long time-courses of pairs of brain regions (Cabral et al., 2017). Due to the lack of a stimulus onset used as a reference in rs-fMRI, comparisons between simulations and measured data are made over a long time window and use time averaged metrics rather than direct comparisons of the predicted trajectories with the measured timeseires (Cabral et al., 2011, 2017; Kashyap and Keilholz, 2019). Researchers are interested in BNM predictions on a time-point basis because many neural processes observed in fMRI occur over these timescales, such as responses to task stimuli or aberrant responses due to neural pathologies. While previous studies have examined faster processes in BNM's, such as comparing them with multimodal recordings such as EEG data (Schirner et al., 2018), no prior investigations have examined how well short-term trajectories, defined by a series of consecutive fMRI measurements, are being reproduced by current BNMs. To address this gap, we solve for initial conditions relative to an observed trajectory for a given BNM and then compare the synchronized predictions of the simulation with the observed timeseries.

We hypothesize that the BNMs that are better approximation of the underlying dynamical system in whole brain dynamics determined using traditional long-term measures, will evolve more closely to the measured rs-fMRI trajectories. An agreement of parameterization between the long-term metric and the investigated short term metrics would support the evidence that BNM's are simulating meaningful trajectories. Moreover with these initial conditions, BNMs can be later used to simulate and investigate neural processes during these short timeframes such as relation to task fMRI behavioral variables.

The initial conditions are estimated, by utilizing a novel method developed in the Machine Learning community that utilizes a sequence of observations and a given dynamical system to output the initial conditions of the dynamical system that would be the closest fit to the current observed data trajectory. The technique, known as Neural Ordinary Differential Equations (ODE), uses a recurrent neural network (RNN) that keeps track of information from previous timepoints, in order to predict the initial conditions of a given dynamical system based on previous observations (Chen et al., 2019). The neural network model is trained via one step prediction, namely from the estimated initial conditions we integrate the known dynamical system to predict the next timestep and compare it with the true next step. The algorithm, therefore, regardless of the dynamical system, gives similar predictions over the first-time interval but the trajectories diverge over longer periods of integration due to differences in the dynamical system and become less dependent on initial RNN predictions.

A potential issue to this approach is that both the signal simulated as well as the measured rs-fMRI signal are thought to be produced by stochastic processes. For simulations this is achieved by adding noise to the models differential equations. This noise might affect the approaches' ability to discriminate correctly between different BNMs on their ability to simulate rs-fMRI, as it adds variance to the data. However, previous studies have shown that despite their variability dynamic metrics are better than metrics computed over a long period of time to parameterize BNM, thus suggesting that even in shorter windows allows for discrimination between models (Kashyap and Keilholz, 2019).

To test whether this approach can correctly identify components of a known BNM, we used the Firing Rate Model (FRM) from Cabral et al. (2012), as a candidate dynamical system to fit to the rs-fMRI data. The FRM is a linear model that defines the change in dynamics in a single neural population as a weighted sum of its network neighbors and applies an exponential decay term to prevent runaway excitation (Cabral et al., 2012). The model contains three components (global coupling, noise amplitude, structural matrix), which are varied independently, and a specific Neural ODE is trained for each variation to solve for the initial conditions. The results show that without noise, maximizing accuracy over the short time window yields trivial BNMs that do not depend on the structural connectivity. However, in the presence of noise the trend reverses and models with strong structural connectivity perform better than the models with weak or no network influence. Since the value of noise is unknown, an additional parameter, namely the structural connectivity was varied by slowly adding noise to the original connectivity, and the sensitivity due to this change in network was measured. The FRM that exhibited the greatest changes in accuracy due to perturbations of the structural connectivity had a parameterization that was in agreement to the one established using long term metrics.

In short, the manuscript demonstrates that Neural ODE approach can simulate FRM trajectories that can be meaningfully compared with measured rs-fMRI data. The differences in parameterizations of the model with respect to rs-fMRI data observed during this timeframe are similar to those observed over longer simulations. Therefore, this tool can be used in the future to analyze BNM on shorter timescales with respect to measured data such as for task fMRI. Furthermore, it can serve as an unbiased metric to directly compare the signal with the models and aid in the development of discovering more powerful models that recapitulate whole brain activity.

2. Methods

2.1. Overview

This section is organized by first describing functions that are used to fit to the rs-fMRI data, mainly BNMs but also certain null models

that are used for comparison. The subsequent sections then describe how the Neural ODE algorithm is used to infer the initial conditions of a given dynamical system based on previous measurements. We describe our own implementation of the Neural ODE algorithm that was specifically designed to train on large amounts of imaging data (Chen et al., 2019). The algorithm was validated using synthetic data from a simple spiral dynamical system described in detail in the Supplementary sections. The subsequent sections after describing the algorithm, deal with the processing of experimental fMRI and DTI data used to construct the models. The final section outlines how the simulated trajectories are compared with empirical trajectories.

2.2. Brain Network Models

Brain Network Models are used as models for whole brain network activity. BNMs combine a mathematical description of the intrinsic activity of a neural population with the global brain structure that coordinates the activity between populations. To construct a BNM, researchers first define a structural network, based on a parcellation scheme that outlines which cortical areas work cohesively as a neural population. In this manuscript, the Desikan Killiany atlas is used as a parcellation scheme as it has been used successfully before for whole brain simulations (Cabral et al., 2011, 2012). Only the cortical areas without the insula are represented in the model constituting a total of 33 regions for each hemisphere for a total of 66 brain regions. These regions serve as the nodes in the network model, while network neighbors are defined using tractography to map out fibers that connect two regions of interest. The change of the activity in the ith brain region is defined as follows:

$$\dot{x_I} = \sum_{J \in Neighbors\ of\ I} F\left(x_I, x_J, \rho_{IJ}\right) + \sum_{K \in \text{Task inputs}} G\left(u_K, \grave{A}_{KI}\right) + \mathrm{N}\left(0, \tilde{A}\right) \quad (1)$$

The first term represents the network component which is described by a function F that depends on its own activity x_i, activity in its neighbor x_j, and the physical properties of the fiber represented by the vector ρ (i.e., the number of fibers between regions, the delay in propagation). The second term consists of a function G that represents external input, whose activity is represented by a k-dimensional vector u representing all sub-cortical and sensory inputs, and the vector π representing again the physical properties that project these inputs into the cortical model (i.e., thalamic tracts into cortex). The last term represents a zero-mean Gaussian noise from the neuronal populations or from omitted higher order terms from the network equations. For resting state activity, the assumption is that $u_k(t) = 0\ \forall\ t$ and the first term dominates the change in activity. This still leaves a large family of functions that are used to approximate F, with many parameters that can widely change the dynamics of the system. In theory, all of these functions can be used to fit the fMRI data with the Neural ODE algorithm, and for each of them initial conditions can be estimated from empirical data. In this manuscript, we focus on the Firing Rate Model, the simplest model that can recapitulate whole brain activity, and use it to solve for initial conditions in the Neural ODE.

2.2.1. Firing Rate Model

The FRM represents the activity of a brain region as the mean firing rate. The change in firing rate of a region depends on a weighted sum of all its neighbors' activity (Eq. 1). The FRM has two parameters: the global coupling parameter k, which controls the strength of network input, and the level of noise amplitude σ, which simulates random activations of brain regions due to unknown neuronal activity (Cabral et al., 2012). At values of $k < 1/(\text{max eigenvalue of } W)$, the system is stable and the system decays to the origin without extraneous noise input. Typical values are $k = 0.9/(\text{max eigenvalue of } W)$ and $\sigma = 0.3$ (based on the Desikan Killiany atlas) where there is a trade-off injecting noise in order to perturb the dynamics and the relative strength of the network to keep the neural areas functionally linked over time (Cabral et al., 2012).

$$\dot{x_I} = -x_I + k \sum_{J \in Neighbors\ of\ I} W_{IJ} * x_J + N\left(0, \sigma\right) \quad (2)$$

2.2.2. Parameter and structural perturbations

Each of the components of the FRM are varied. These components include the global coupling parameter k, the amplitude of the noise level σ, and the structural weight matrix W. The global coupling parameters and the structural matrix are fixed before using the Neural ODE algorithm, such that for each specific set of values for k and W a separate LSTM network is trained to generate initial conditions. To vary the global coupling parameter, the parameter k in Eq. 1, is adjusted from 0.9 to 0 in 0.15 intervals. To generate structural perturbations, a random percentage of the original edges are swapped to connect two different nodes while keeping the graph symmetric. This creates random perturbations from the original structural matrix while maintaining the number of edges. Each of these graphs would result in different dynamics, but the trajectories from the model containing the original structural connectivity should be the closest to the measured rs-fMRI data. The noise parameter is not used in training the LSTM Neural ODE model. However, during testing after the initial conditions are estimated the noise is introduced by testing σ using values [0.0001, 0.15, 0.3, 0.45].

2.2.3. Null models

To compare the effects of fitting to the Neural ODE with other functions, the rs-fMRI data is fitted with null models that do not simulating network activity. The simplest of these models sets the Neural ODE equations to 0, such that the prediction of the LSTM is the output of the model. This quantifies how well the LSTM network's initial condition prediction matches the next predicted output without any of the BNM functions. For future timesteps, it acts as a simplified autoregressive model by holding the current input as the output, i.e., $x(n+1) = x(n)$. This is implemented by setting the connectivity matrix in Eq. 2 to an identity matrix which cancels out with the first term and sets the equation to 0, and is referend to as the Autoregressive (AR) model. The second null model is obtained by setting the global coupling parameter to zero in the FRM and the differential equations reduce to an exponential decay. This model should perform worse than the BNM equations but test the limits of the global coupling values. Finally, we compare it to a pure Machine Learning Inference model as published in Kashyap and Keilholz (2020), where at each

timestep, the output of the LSTM is fed in as the next input. This model is non-deterministic as the output of the LSTM is sampled from a distribution. The function implemented by the LSTM in this case is completely unknown, as even the noise level changes as a function of the input. This model, however, gives a good estimate of an upper-bound on predictability on a short time window and does better than using traditional BNMs and can replicate complex resting state processes as Quasi Periodic Patterns and dynamic functional connectivity analysis performed using K-means (Kashyap and Keilholz, 2020).

2.3. Neural ordinary differential equations

The Neural ODE algorithm is designed to estimate the initial conditions of a given dynamical system based on past observations. Figure 1 provides an overview of the algorithm, which involves fitting to a spiral dynamical system based on noisy observations using Neural ODE. The task of the recurrent neural network is to predict the true initial conditions of the spiral dataset (shown as blue underlying trajectory in Figure 1) based on the sequence of observed measurements shown in green. A RNN implementation known as Long Short Term Memory (LSTM) was used to perform this task, as it keeps the information of past data observations [x_0 to x_{t-1}] in its hidden state p_{t-1} (Graves and Schmidhuber, 2008). Thus, when the timeseries is fed into the RNN one timepoint at a time, the current information is incorporated into the hidden state and is passed forward as shown in the LSTM unrolled version, in order to aid in the prediction of future observations. The LSTM's predictions become more accurate as it observes more data, up to a certain limit beyond which newer data does not add any new information to the hidden state (Graves and Schmidhuber, 2008). For this particular task, the LSTM's output is defined as the initial conditions of a given dynamical system. Since the initial conditions are not known and thus an effective gradient cannot be computed based on initial conditions alone. Therefore, the algorithm assumes that the next observation is the integral of the predicted initial conditions and the given dynamical system with some noise added to it (Chen et al., 2019). The loss function is calculated based on the measurement at the next timestep in ensure the output of the LSTM to converges to the correct initial conditions for the given timepoint. A schematic of the algorithm as well as the Tensorflow implementation are shown in Supplementary sections 7.1, 7.2.

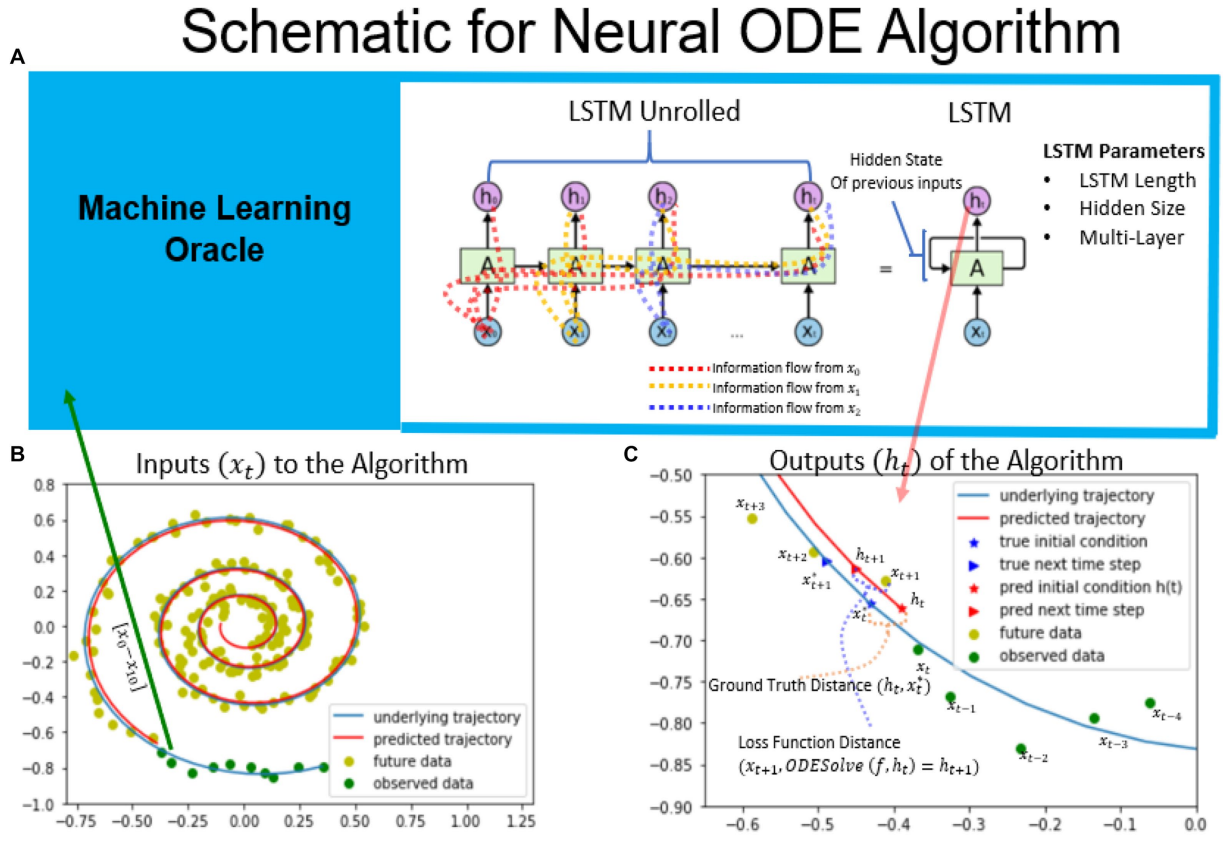


FIGURE 1
Schematic of the neural ODE algorithm. Schematic of the neural ODE algorithm. An example spiral trajectory is shown in panel **B** with green points representing the data sequences. The RNN takes in one data point at a time and updates its hidden state as well as outputs its prediction for the initial condition. The hidden state keeps track of information from previous observations and is carried forward to future time steps, as illustrated in the RNN unrolled diagram. The output of the RNN represents the initial condition of the dynamical system at that timestep. For the spiral dataset, the true ground truth initial condition x_t^* is known and is illustrated in bottom right. The ground truth distance is used to evaluate the trained RNN network's ability to predict the initial conditions in the constructed dataset. The next timepoint is predicted by integrating the ODE system based on the given dynamical system. The loss function is defined as the difference between the next predicted timepoint and the next observed time point, x_{t+1}. Minimizing the loss function distance minimizes the distance to the ground truth initial conditions. (Mars, 2020).

2.4. Experimental data

2.4.1. Structural network for Brain Network Model

To estimate the structural network, tractography was run on 5 HCP Diffusion Weighted Images using the freely available software Mrtrix (Van Essen et al., 2013; Kashyap and Keilholz, 2019). The fiber orientations in the DWI images were first estimated using constrained spherical deconvolution. Next, using a probabilistic streamline algorithm, 100 million fibers set at a maximum length of 250 mm were computed for each individual and then filtered to 10 million fibers. To construct the structural network, we determined the number of fibers that intersected two ROIs in the Desikan-Killiany atlas and normalized the power by dividing by the surface area of the receiving region (Desikan et al., 2006; Hagmann et al., 2008; Cabral et al., 2011). Finally, the matrix is normalized by dividing by the largest eigenvalue such that the graph Laplacian (k*SN-I) has only negative eigenvalues (Cabral et al., 2012). This normalization ensures that the feedback decays and prevents an exponential increase in the signal over time.

2.4.2. fMRI data

The fMRI data used to test and train the models were obtained from the Human Connectome Project 447 Young Adult subjects release. The scans were pre-registered to Montreal Neurological Institute (MNI) space in surface format (MSMAII) and denoised using 300 Independent Component Analysis, following the recommended steps by Salimi-Khorshidi et al. (2014). The surface-vertex or grayordinates time series were transformed to the ROI time series by averaging all vertices based on the Desikan-Killiany atlas parcellation. This was done on an individual level since the surface parcellations are provided to by HCP and Freesurfer for each individual subject (aparc and aprac2009 files). The signal was then bandpass filtered from 0.0008 to 0.125 Hz and then the global signal was regressed using a general linear model using the mean timeseries of all cortical parcels as the global signal. Finally, the signal is subsequently z-scored as described in Kashyap and Keilholz (2019). For the task data, each dataset (language, working memory, motor, social, emotional, gambling, relational) was processed separately and then concatenated together. Each task dataset was truncated to the closest multiple of 50 timepoints and the autoencoder was fed alternating segments of task and the rest data for training. The algorithm was trained using both task data as well as rest data, because the algorithm performed better on most metrics with more varied data. Furthermore, it is believed that during task activity, resting state networks dominate most of the cortical activity and task networks often look indistinguishable from rs-fMRI networks (Smith et al., 2009). However, during evaluation, only the results on predicting future rs-fMRI were presented, while task will be addressed in future work.

2.5. Metrics and evaluation between simulated and empirical trajectories

The dynamical models are evaluated on how well they fit with the empirical observations from the estimated initial conditions. For the spiral dataset, the true initial conditions were known, allowing for direct calculation of results using a Euclidean distance between the estimated and true initial conditions. For the fMRI data, the r-squared and the mean squared error at each timepoint between the predicted and observed data vectors representing the activity of 66 brain regions were calculated. Since the loss function of the Neural ODE algorithm, converges to zero during training across most models, this metric tends to be most similar when comparing across models (see Supplementary section 7.7). Therefore, in order to differentiate between the models, the error was calculated for subsequent timepoints to gauge how well the trajectory follows the timeseries over a longer period of time. The results were calculated across a set aside test dataset using a batch of subjects ($N = 80$). The Supplementary section 7.5 discusses the differences in variances between group and individual models, and the effect of testing at every timepoint versus a few timepoints. While there is no pronounced difference in testing a few timepoints and evaluating the system at every timepoint, there is a large difference in variance between the averaging the error out in a batch of subjects ($N = 80$) vs. in the individual models. The group metric was selected for its smaller variance and greater robustness, given the purpose was fit a group model rather than an individual model for the HCP resting state dataset using the Neural ODE algorithm.

3. Results

The objective of this study was to assess the feasibility of using Neural ODEs to solve for the initial conditions of BNMs, and subsequently differentiate between different models based on how well they follow the resulting rs-fMRI measurements. To validate this approach, the methodology was applied on a constructed spiral dataset where the true underlying dynamical system was known and the correct coefficients are shown to be determined using this approach. Subsequently, the technique is applied to fit with the rs-fMRI dataset using a FRM, and the model parameters and coefficients are estimated and shown to be similar to previous literature.

3.1. Differentiating between dynamical systems on spiral data

The process of generating the spiral dataset was explained in detail in Supplementary sections 7.3–7.7, but, in essence, it involved a two variable linear dynamical system and that was often used as an example in machine learning literature to show the feasibility of solving for initial conditions. The Supplementary sections began by showcasing the results from the previous paper Chen et al. (2019), where the Neural ODE algorithm's predictions converged to the right initial conditions after observing a sufficient number of previous timesteps to allow the LSTM to make valid predictions. The subsequent sections detailed the methodology used to determine the network hyperparameters such as the hidden size and number of layers. Larger networks were found to be more sample-efficient as they could predict the initial conditions based on fewer timepoints. At a certain size, the accuracy did not improve much with alterations to the network and was used as the model to perform the following experiment on system identification.

The spiral data was generated using a known system of differential equations as presented in Figure 2 (top right). The objective of the Neural ODE algorithm was to solve for the initial conditions for each

of the candidate dynamical systems with respect to the observed data. Each of the systems contained distinct coefficients in their respective weight matrices, where W1 represented the original dynamical system and W2, and W3 contained the original structural matrix perturbed with increasing noise. The findings illustrated in Figure 2 depicted three instances of fitting these spirals to generated spiral data for different weight matrices. The results demonstrated that, in short time periods after the initial conditions, the other candidate systems fit the data as well as the original system for the first few points after the initial conditions. Nonetheless, in the long run, the spiral matrix W1 was found to be the closest to the data in Euclidean distance, as shown in Figure 2 (bottom right). Therefore, since the Neural ODE fits any dynamical system tangentially in time, it is important to observe dynamics sufficiently long enough to differentiate between the models. At very long intervals, the distance starts to decrease as all trajectories converge to an attractor based at the origin, which is special for the spiral dynamical system and not present in the neural data. In summary, the results on the synthetic spiral data show that it is possible to use this method as a system identification, but the systems need to be simulated for a long enough time interval for the differences to manifest, as the distances close to the initial condition are harder to tell apart, since the output of the LSTM minimizes the prediction error at the first timestep.

3.2. Fitting differently parameterized firing rate models to resting state fMRI data

The dynamics of BNMs were influenced by many parameters and were tuned to fit fMRI data. Therefore, it was essential to test whether this method allowed us to parameterize different BNMs. The FRM was chosen as it has been well studied in the past, and the Neural ODE can be validated by reproducing previous estimates (Cabral et al., 2012).

Figure 3 compares various differently parameterized FRM models and three Machine Learning Null models in terms of their ability to reproduce the future trajectory. After estimating the initial conditions,

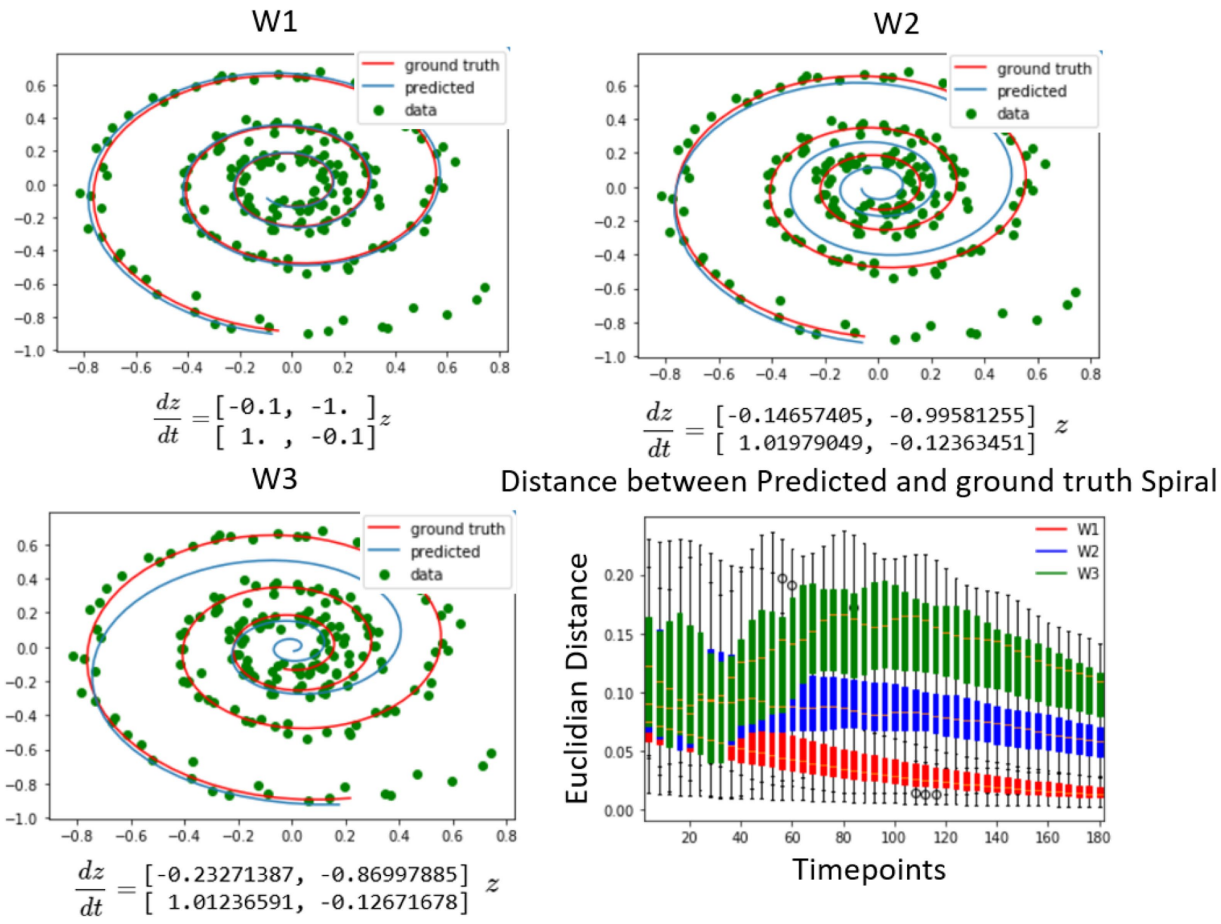


FIGURE 2
Differentiating between dynamical systems on spiral data. The study utilized a number of candidate spirals, which were a perturbed version of the ground truth spiral, to fit using the neural ODE algorithm to the spiral data. The top left shows a spiral that was well aligned with the data, as it was used to generate the data. The top right and bottom left figures showed spirals that were increasingly further away from the ground truth. The neural ODE algorithm was able to fit any of the spirals to the data for a given set of observations, but over time, the candidate spirals that were further away from the ground truth diverged much faster from the future data points. This divergence was quantified by the plot on the bottom right, where the distribution of the distance between the three different spirals and the observed data points was plotted from the predicted initial conditions. Initially, the distance between the spirals of different weights and the observed data points was close, but it diverged further away when compared to future timepoints. It is noteworthy that due to all trajectories going toward the origin, the distance at very large timescales converged to zero, but this is not expected in the brain data, where the signal does not approach a single attractor. (Mars, 2020).

the distribution of distances between the predicted and the actual trajectory was plotted over time. At the first timestep (Figure 3, top left), all the models perform relatively similarly as a direct result of minimizing the loss function using an LSTM. Similar to the spiral example, the models diverge in performance when moving forward in time. Surprisingly, at the fourth timestep (Figure 3 top right), the exponentially decaying null models with a zero global coupling, and the autoregressive null model without a BNM (labeled as AR) perform better than models that contain the brain structure. This suggests that introducing any BNM, decreases the accuracy of the model, and the model performs best using just the LSTM predictions. The null model utilizing LSTM inference (Kashyap and Keilholz, 2020) performs the best and represents an estimate of an upper bound in predictability of the rs-fMRI signal.

However, interestingly this trend completely reversed in the presence of noise. In Figure 4, both the standard deviation of the noise as well as the global coupling parameter have been varied. The models with low global coupling perform better at low noise levels, but as the noise level increases, the BNMs with stronger network effects outperform those with low levels of global coupling. This suggests that noise plays a critical role in establishing the parameters of the FRM, and the properties of the structural network become more significant when the system has high noise. The overall r-squared of the models decreases with the introduction of noise, but the rate at which they diverge from the measured trajectories appears to depend on the global coupling parameters. Previous FRMs that used the same brain parcellation, were simulated with $k = 0.9$ and $\sigma = 0.3$, and noise was seen as essential in simulating the BNMs (Cabral et al., 2012). However, since both parameters are unknown and the overall r-squared decreases with the introduction of noise, just based on varying these two parameters it is difficult to conclude which parameterization yields the best result using this approach. Moreover, the AR null model still performs better than the introduction of a BNM, but the difference is much smaller than before. The inference model is not included in the noise

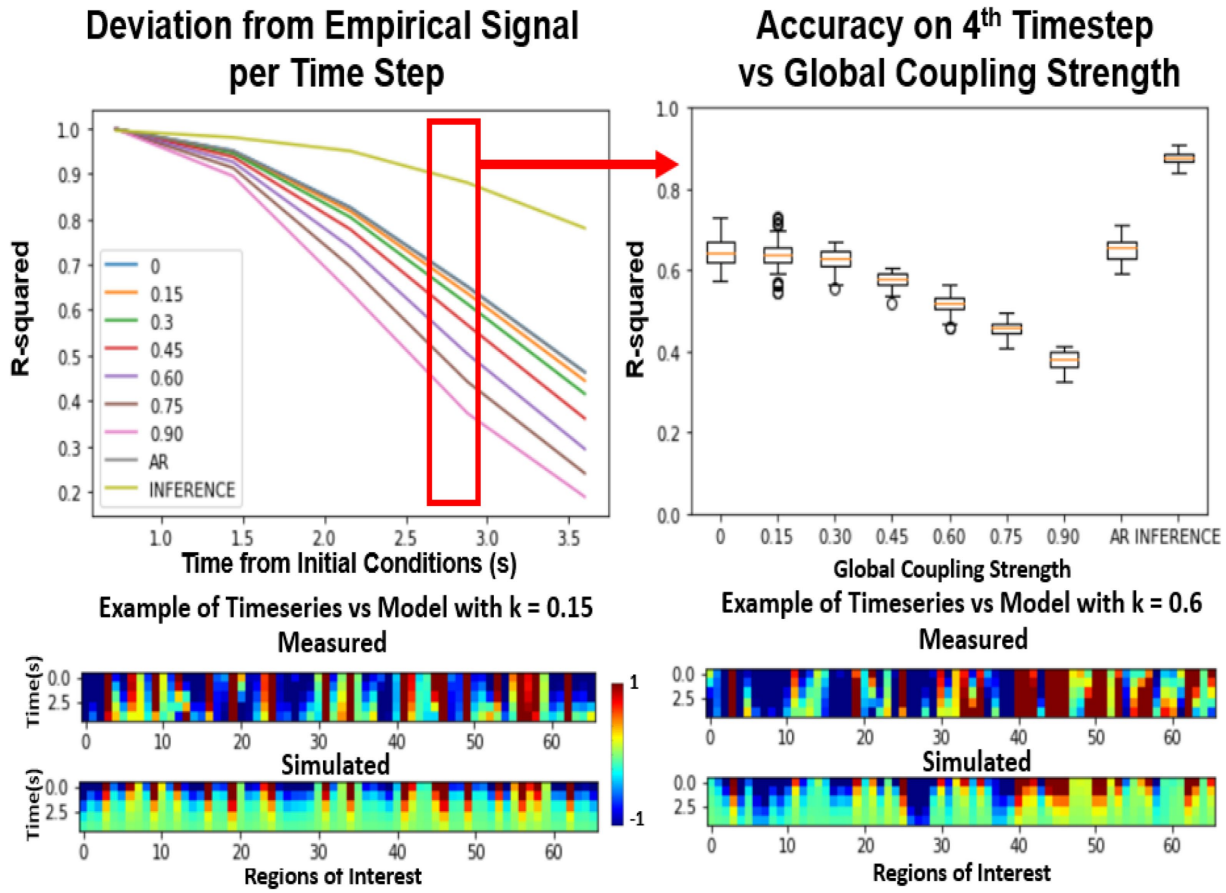


FIGURE 3
Effects of global coupling in the noiseless firing rate model. Evaluating the FRM (Eq. 1) with different global coupling parameters. The performance was quantified by the r-squared value between the simulation and the model. Top Left: Error per timestep from the estimated initial conditions for various different parameters compared. Examples of the model timeseries vs. the resting state timeseries are given on the bottom at two different parameterizations but the distribution shown in the top panels quantify their performance across 2500 trials across unseen test data. Top Right: At the fourth timestep, the distribution of the r-squared across all the models is plotted. For the FRMs, the accuracy decreases with the increase of global coupling, and the model with zero global coupling performs the best. The Autoregressive (AR) model utilizes the LSTM for the first timestep prediction and then outputs the next prediction as the previous timestep and does as well as the FRMs with zero global coupling. The inference model using LSTM at every timestep as implemented in Kashyap and Keilholz (2020) performs the best in terms of accuracy, but the model dynamics are unknown as they are implemented using deep learning. (Mars, 2020).

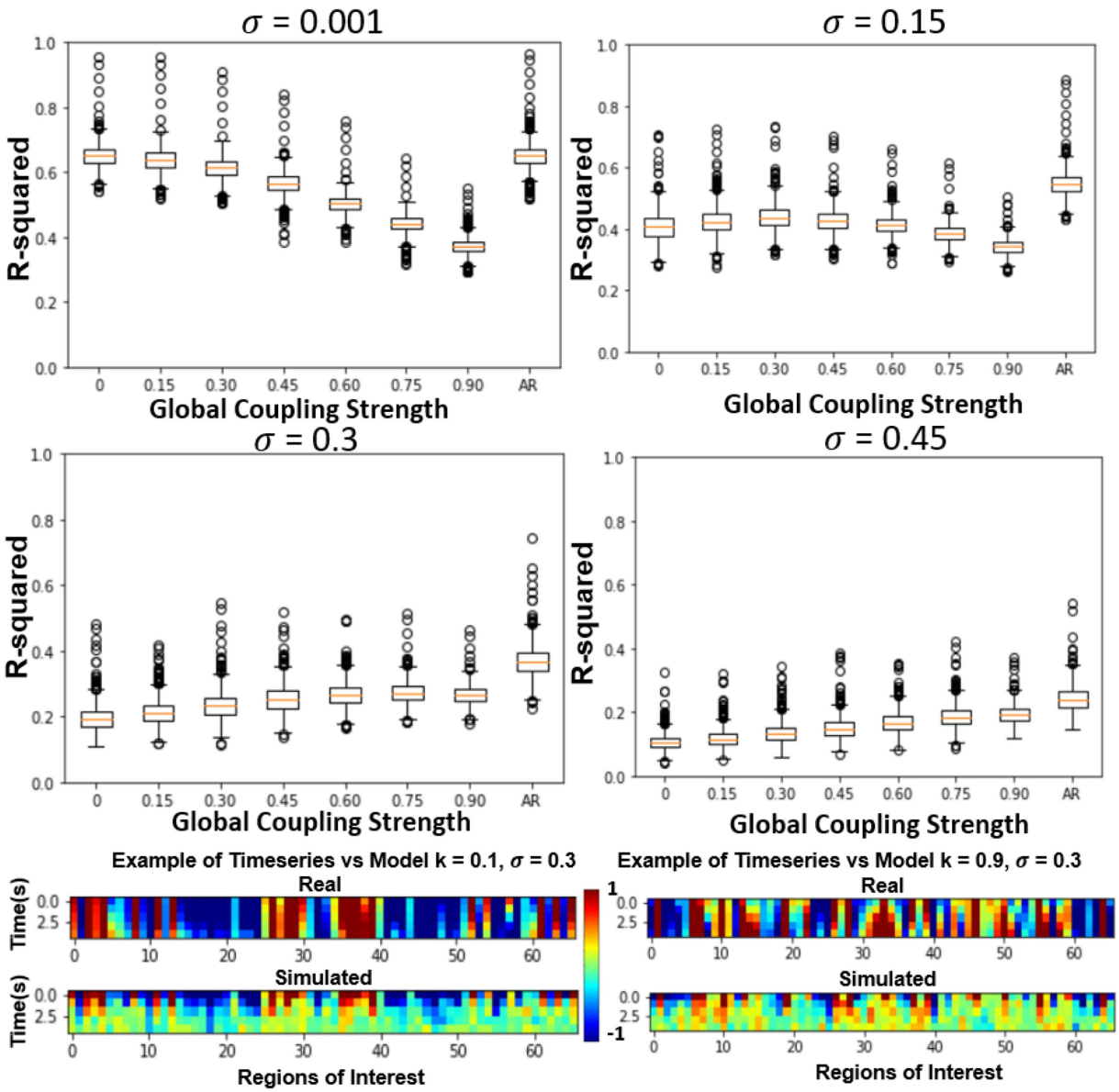


FIGURE 4
Effect of varying global coupling and noise on the 4th timestep accuracy of the firing rate model. The effect of varying the two parameters in the FRM, the global coupling K and the standard deviation of the noise σ. The introduction of noise lowers the accuracy of all models but does so in an uneven fashion. At low levels of noise, the exponential only model (k=0) outperforms the BNMs with structural connectivity matrices. However, with increasing noise power, the BNM with stronger global connectivity matrices appears to outperform the naïve exponential models. The autoregressive model's performance also worsens with increasing noise levels, although it still outperforms the FRMs regardless of the coupling strength, the gap between them is reduced. (Mars, 2020).

estimations, as the dynamical system is represented as a RNN and cannot be manipulated in a controlled manner as the other models.

The introduction of noise changes the resulting dynamics, which is apparent from the time traces. Illustrated in bottom of Figure 4, the trajectories without noise decay to zero, in line with well-known analytical solutions to the consensus equation, where the values of a connected network with eigenvalues less than 1 converge to the origin (Mesbahi and Egerstedt, 2010). However, the introduction of noise results in more complex trajectories as depicted in Figure 5 bottom, where the values do not decay to zero, but rather randomly oscillate around the origin which serves as an attractor in the system (Cabral et al., 2011). The role of the structural network, in this case becomes more important as it integrates the noise inputs through the network, and results in trajectories more similar to the measured rs-fMRI signal.

3.3. Differentiating between BNM due to differences in structural connectivity

In the previous section, only the parameters of the FRM, the global coupling strength as well as the magnitude of the noise were changed. However, in this section, the effects of simulating six different SC matrices at high and low different global coupling

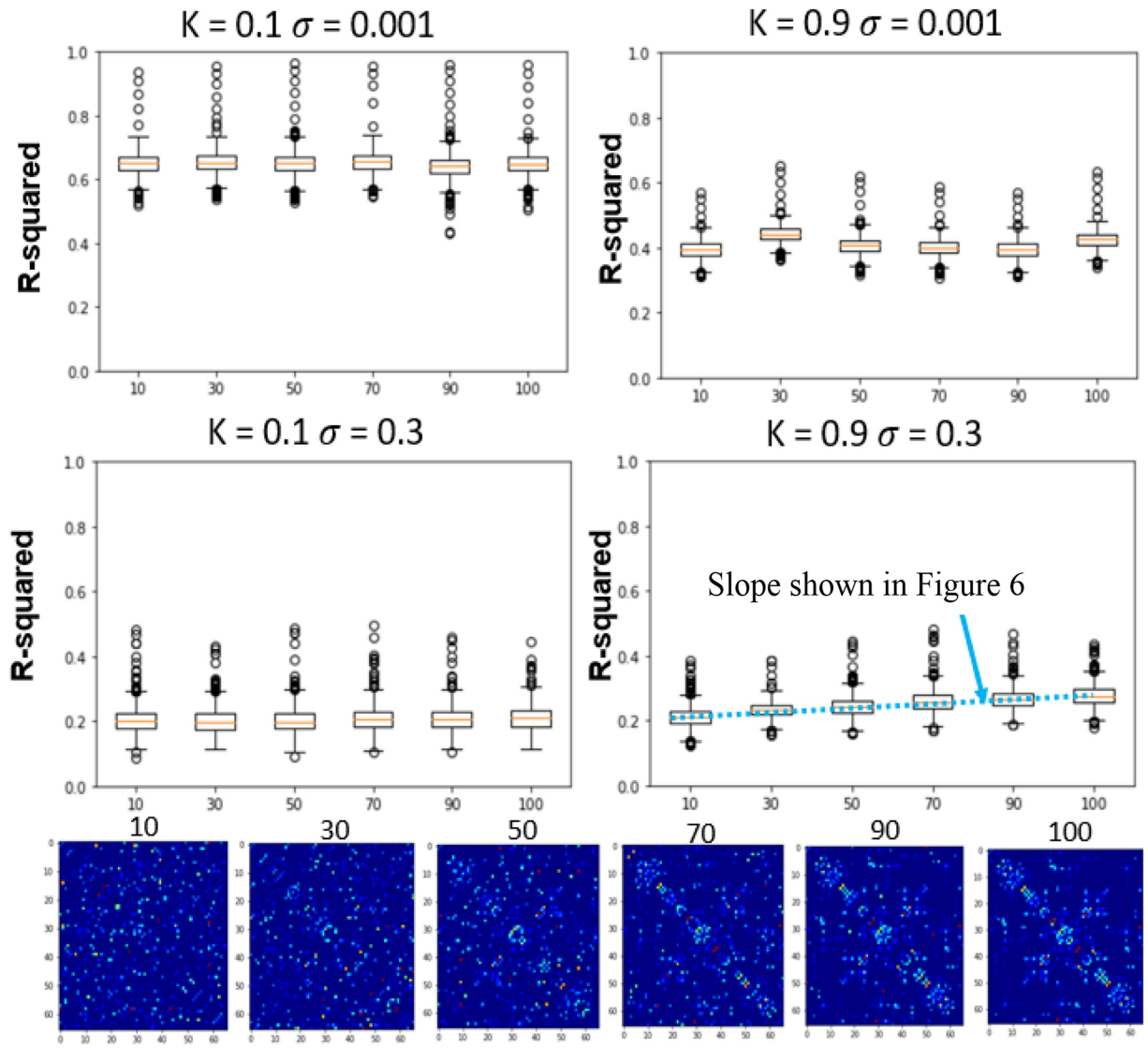


FIGURE 5
Effects of global coupling, noise, and structural connectivity on the 4th timestep accuracy of the firing rate model. Examining the effects of parameterization and estimating the correct SC. At top left, we show the results of changing the structural connectivity for a low global coupling model and low noise levels. It does not vary as a function of the structure and performs relatively similarly to the LSTM only null model. At high global coupling and high noise levels, the models show that they are more of a function of the correct structural network (bottom right). The slope across the performance of different structural connectivity (bottom right of Figure 5) is used as a metric in the next section to solve for global coupling and noise levels. (Mars, 2020).

($k = 0.1, 0.9$), are quantified while varying the high and low noise levels ($\sigma = 0.001, \sigma = 0.3$). The SC matrices are varied from the measured SC by flipping edges and results in SC seen in Figure 5 bottom row. The r-squared value at the fourth timestep, between the different models is plotted in Figure 5 top two rows. Unlike the previous sections where the correct parameter values were unknown, in this experiment, the original SC is expected to outperform the models with altered SC configurations.

At low noise levels ($\sigma = 0.001$), there was no significant relationship between altering the structural connectivity and either of the coupling strengths. However, at the high noise levels ($\sigma = 0.3$), although the model has a lower r-squared than at the low noise levels, the effects of the network are evident, with the original SC configuration outperforming the corrupted SC configurations for both low and high coupling strengths. The trend was once again more prominent for the high global coupling ($k = 0.9$) than the low global coupling ($k = 0.1$).

3.4. Estimating the parameterization and noise level of the firing rate model

In the previous sections, the effects of varying the global coupling and noise levels on the accuracy of the simulated system were explored, but the relationship between the system and the underlying structural connectivity remained unclear. To investigate this, the system was simulated with different SC matrices while varying the noise levels and global coupling values. The slope was calculated at different time steps to find where the system was most sensitive to the

underlying structural connectivity to find the correct parameterization of the models. The slope is plotted from timesteps 2 to 5, across different global coupling values and noise steps in Figure 6. At timesteps close to the initial condition, higher levels of noise are needed to differentiate the systems sensitivity to the structural matrix. However, at later timesteps, the opposite is true where higher noise levels perturbs the system too much and the overall r-squared drops so low that the models become indistinguishable to each other. Therefore, the fourth timestep is used where the max differentiation between models occurs regardless of the coupling values, where the trajectory is far enough from the effects of the LSTM fitting and close enough in time to test the predictability of the models. At this timestep, the maximum occurs at the values ($k=0.925$, $\sigma=0.35$). This value is very close to what has been used to be simulate FRMs ($k=0.9$, $\sigma=0.3$) from previous publications (Cabral et al., 2012). Their approach of parameterization here has been reproduced in Figure 6 (right most panel), which calculates the FC of a 20 min simulation and correlates the FC with the FC of the empirical data. The maximum at [0.875, 0.3] is in good agreement with the short-term measures and the previous estimates.

3.5. Evaluating the initial conditions of the NODE model

In the previous sections, the application of NODE algorithm to correctly bias the BNM models and recover coefficients by using short term metrics that match those of previous literature were highlighted. In the following section the NODE ($k=0.925$, $\sigma=0.35$) is utlized to evaluate the initial conditions of the algorithm vs. null initial conditions. The null initial conditions were generated by taking the previous timestep, and integrating the BNM from that timestep. For long term simulations shown in Figure 7A, the functional connectivity is characterized in Figure 7B with a correlation of 0.45 with the empirical measured signal. While the initial conditions did not change the functional connectivity of a long term simulation, Figure 7C illustrated that the trajectories from NODE initial conditions followed the signal more closely than the null initial conditions. Moreover, this difference is also present when comparing rest vs. task as shown in Figure 7D , indicating that the algorithm is producing non-trivial results for its initial condition prediction.

4. Discussion

4.1. Overall discussion and significance

The study proposed the use of the Neural ODE technique for estimate initial conditions in different candidate BNMs and subsequently evaluating the predicted trajectories compared to the real data. To test this methodology, the technique was first applied to a well-studied spiral dataset, which demonstrated its ability to correctly identify parameters in a constructed example of system identification where the ground truth was known. The method was then applied to fit different Firing Rate BNMs to neural fMRI data by varying their parameterizations, noise level, and by changing the structural connectivity. By using all three, the system was correctly able to identify the parameters of the FRM, which were close to previous estimations of the model's parameterization using the same whole brain parcellation (Cabral et al., 2012). Moreover, these initial conditions were shown to be non-trivial as they perform better than just using the previous timepoint as initial conditions.

Therefore, pertinent information to parameterize BNMs is present in short term trajectories analysis. Unlike older metrics, this technique allows for direct timeseries comparisons between the theoretical models with measured experimental data, and circumvents the reliance on a certain metric/interpretation of rs-fMRI. Therefore, this technique provides a unbiased metric that can be extended to compare and parameterize more complex BNMs. Moreover, it allows pathway forward in studying whole brain dynamics on a faster timescale, and illuminate what our current theoretical models can and cannot explain in terms of transient dynamics.

The interpretation of the initial condition is difficult in rs-fMRI, as rest is not labeled with respect to stimulus, but they can provide

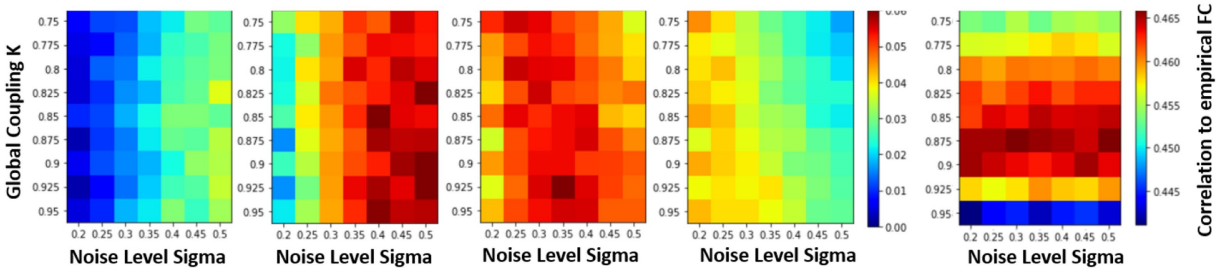


FIGURE 6
Parametrization of FRM using short term measures vs. parameterization using long term measures. Examining the effects of changing the structural connectivity matrices vs. accuracy (slope in Figure 6) under different parameterization/noise levels of the Firing Rate Model. The sensitivity of the system to changes in the SC matrix is the only metric where the expected outcome is observed, where the original structural matrix outperforms a random one. The change in r-squared value is represented using a color bar and is plotted for different timesteps. At the fourth timestep ($t=2.88$) the maximum differences occurs on all the models regardless of the parameters, since at that time step the system has diverged enough from the initial conditions and while not far enough in time to cause the overall r-squared to drop too low and make the models indistinguishable. At this timestep, a maximum slope occurs at ($k=0.925$, $\sigma=0.35$). Comparison to the traditional parameterization is shown on the right. Here the differently parameterized FRM models were simulated for 20min and then the FC matrices of the resulting simulation were compared against the empirical FC. The traditional approach has a maximum at [0.875, 0.3], and previous reproductions of this experiment found a maximum at [0.9, 0.3] (Cabral et al., 2012), showing good agreement on which BNM recapitulates rs-fMRI in both short and long term measures. (Mars, 2020).

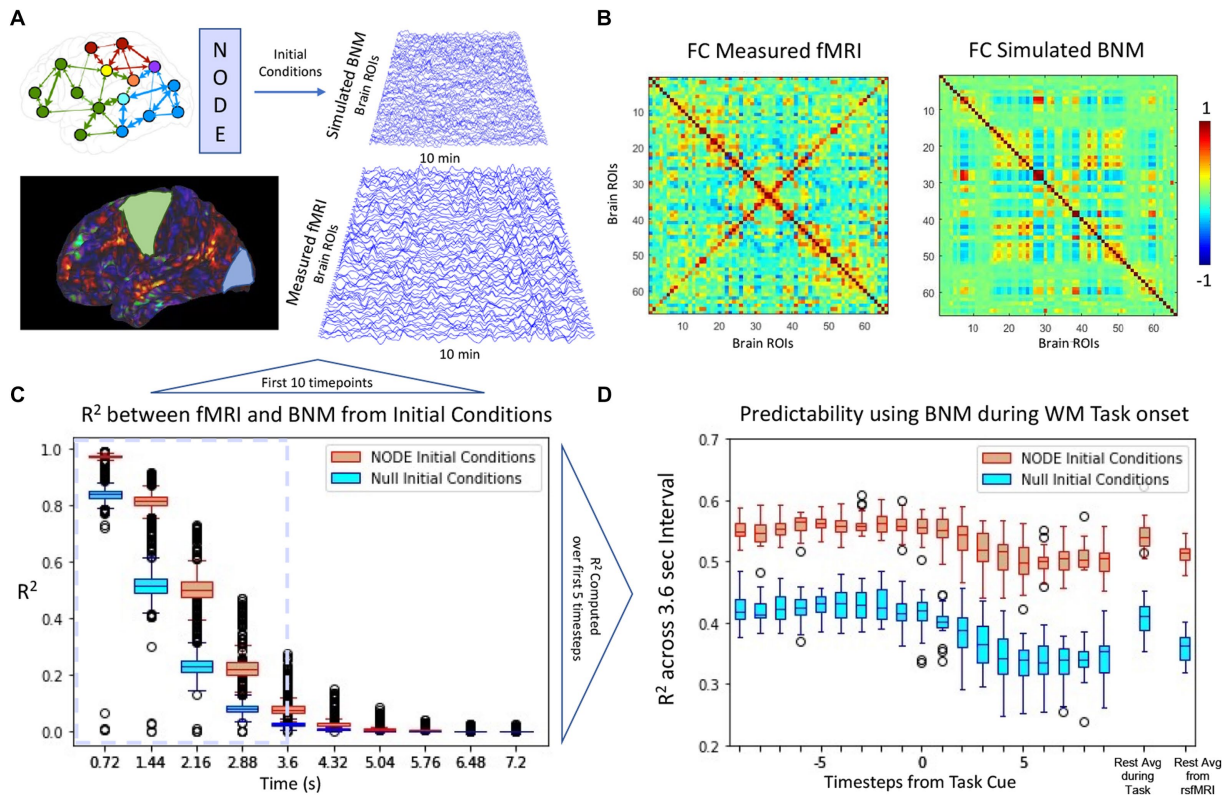


FIGURE 7
(A) Comparison of BNM simulation with measured brain activity derived from rsfMRI. Simulation of 66 regions for 10min is first synchronized using the NODE algorithm such that the initial conditions are set for the BNM. **(B)** Functional connectivity of the simulated model and the empirical signal (cross-correlation ~0.45). **(C)** For the first few timepoints, the simulated signal follows trajectory more closely using the solved for initial conditions than using the measurement at that timepoint as initial conditions (null initial conditions). The signals diverge due to the noise added at the simulation as well as the drift that occurs in rsfMRI. We define the Region of Predictability (RP) based on the first 3.6s and compute the across all timepoints in that interval, at which the NODE initial conditions perform significantly better than the null model. **(D)** The RP is plotted from every timepoint during a working memory task, where the cue is presented at 0. All models lose predictability during the task onset and predictability stays lower during the task interval compared to the average rest average. The predictability during the task onset is closer to the predictability during the rsfMRI scans than during the resting state portion of the task scans. (Mars, 2020).

information of the phase of cyclical brain processes such as quasi periodic patterns (QPPs) that have been identified in the literature to exist in both rest and task data (Thompson et al., 2014). The phase is thought to be important improving the correlates to response time in task fMRI, and thus estimating these conditions might prove relevant in understand the current state of these cyclical brain processes (Abbas et al., 2020).

4.2. Parameterization and noise levels of the firing rate model

The study focused on the Firing Rate Model (FRM), which is the simplest of the BNM used to simulate rs-fMRI. The model has two variable parameters - the magnitude of global coupling and the level of noise (Eq. 1). Previous literature has suggested the global coupling value of a traditional FRM is set slightly less than 1, around 0.9, which is just before the system becomes unstable. The noise is usually set around 0.3 and global coupling to 0.9 for this parcellation scheme (Cabral et al., 2012; Kashyap and Keilholz, 2019), where closer toward zero it simplifies to the well-known consensus problem where the timeseries converge to the attractor at the origin (Mesbahi and Egerstedt, 2010), and at higher degrees of noise the system becomes completely chaotic and non-deterministic. Searching for the correct parameterization of the FRM between the global coupling and the magnitude of the noise is therefore an important to simulate the model in the correct regime.

In practice, this relationship was not so easily ascertained by analyzing the short-term trajectories as it was confounded by the presence/absence of noise. The simulated trajectories that were the closest to the empirical trajectories were models that contained no noise resulting in a parameterization of a trivial exponential decay null models (Figure 3). However, in the presence of noise, the role of network structure became important, as at higher global coupling values the signal would deviate less from the empirical trajectories. The role of the structural connectivity here can be thought to averaging out the noise, and the trajectory became more robust to local deviation due to noise introduced at each ROI. Supporting this argument, Figure 5 demonstrated that in the presence of noise, the models also exhibited a dependence to changes in the structural connectivity, where the true structural connectivity resulted in dynamics closer to the empirical signal than noisy perturbations of the original structural connectivity. Although the exact value of noise cannot be solved by maximizing the r-squared accuracy while varying the noise amplitude as it results in favoring noiseless models; at the right noise/ global coupling parameterization, the accuracy of the

FRM should be maximally dependent on the correct structural connectivity. This hypothesis was tested in Figure 6, where the change in accuracy of due to the changes in the structural connectivity was plotted. Using this metric, the FRM with ($k = 0.925$, $\sigma = 0.35$) is the most dependent to changes in the structural connectivity and is very close to previously known values that was being used for the FRM ($k = 0.9$, $\sigma = 0.3$) as well as our computed maximum using long term FC estimates ($k = 0.875$, $\sigma = 0.3$) (Cabral et al., 2012). The previous process used long term FC as a metric to maximize to parameterize the models, rather than using the short trajectories as in this manuscript, but here they show they give similar estimates on which FRM is closest to measured rs-fMRI dynamics. The evidence that these two values are close, suggests that our approximation of the true underlying dynamical system is at least scale free across the observed timeframes and models that have been used to simulate long periods of time, can capture meaningful dynamics in the shorter timeframe.

4.3. Comparison to other neural ODE architectures

The original Neural ODE implementation uses a backward time architecture, where the timeseries is inverted and fed into the RNN network, such that the first timepoint is fed into the RNN last and the final prediction is used to infer the initial condition of the whole timeseries and then integrated forward in order to compute the loss function (Chen et al., 2019). They do not evaluate the RNN prediction at every timepoint like in our implementation, but explicitly state that such an architecture would speed the training process. The Tensorflow RNN implementation page also recommended a parallel use of the RNN in order to speed up the training process.[1] The innovative backwards time architectural method gets rid of the initialization problem of the RNN that exists in our forward time implementation but runs into a causality problem where future inputs influence the predictions of previous initial condition. Because BNMs are defined as a function of previous network activity, and because our intended use of the trained model is a continuous correction of the accompanying BNM model, the time forward architecture is used in order to solve for the initial conditions. The other significant difference is that our implementation of the Neural ODE also uses a LSTM after the ODE integration (Chen et al., 2019). This methodology is extensively evaluated in Kashyap and Keilholz (2020) but confounds our goal of comparing the fit of different dynamic systems, so it is simply presented as a null model labeled as inference in this paper. This model outperforms all other models in terms of short term prediction, but cannot be manipulated as in terms of the noise level, coupling strength or other meaningful biological variables. Rather it represents an estimate of an upper bound in terms of predictability seen in the rs-fMRI dataset.

4.4. Comparison to other techniques in literature

The Neural ODE algorithm presented here is a relatively new technique first presented in 2019. To our knowledge this exact technique has not been applied in the context of fitting whole brain models with empirical rs-fMRI data. However, our methodology is quite similar to our own previously published work (Kashyap and Keilholz, 2020), but differs in the important following manners. In the previous paper, the system was trained in a very similar manner, but in the generation of new data from the initial conditions the older methods utilized the entire Machine Learning architecture, LSTM and the Brain Network Model to synthesize new data, whereas in this paper, the future timeseries is generated from initial conditions by integrating the Brain Network Model. The older method allowed to generate more realistic brain data and replicate brain dynamics better than traditional BNM as it utilized the LSTM in every timestep. However, this brought into unknowns into the dynamics and it was not possible to evaluate the BNM on their own. Therefore, in order to isolate the performance of BNM for the purposes of system identification, the LSTM was excluded from the inference process and was only used to generate initial conditions. A recent preprint (Wun, 2020) also utilizes the Neural ODE approach to fit to rs-fMRI data. However, in that methodology it does not use the Neural ODE tool to fit trajectories from the BNM rather analyzes latent variables of a model to predict task states. Many other approaches have started using different techniques for uncovering principles of dynamical systems in order to represent rs-fMRI (Zalesky et al., 2014; Hjelm et al., 2018; Vidaurre et al., 2018; Nozari et al., 2020; Singh et al., 2021). Nozari et al. (2020) uses a similar r-squared metric to quantify the difference at the first time point prediction but does not extend this by predicting further out in time. In this paper, to our knowledge is the first to use these tools for comparing short term trajectories of given BNMs to measured rs-fMRI data.

4.5. Assumptions and limitations

The error from the model's prediction comes from multiple different sources such as (1) the mismatch between the differential equations and the actual dynamics, (2) from the error in predicting the initial conditions, and (3) inadequate descriptions of structural connectivity and/or the lack of including subcortical areas in the simulations.

1. A major limitation of this approach is to have an estimation of the underlying dynamical system that represents the data. This requires vast knowledge of what model including the specific parameterization might fit the dataset. However, the Neural ODE system is able to tangentially fit any dynamical system even trivial ones such as the exponential decay. Therefore, it is not really necessary to have a really good estimation of the underlying system and can be tested how well they predict subsequent timesteps.
2. We assume that for any assumed dynamical system the error from the RNN is uniform no matter what the function is, and the subsequent error calculated from the trajectories is due to the mismatch between the data and the dynamical system. However, this might not be true, and more complex models might have a larger errors in estimating initial conditions and therefore is a potential confounder in our analysis.
3. The inadequate description of the brain network is also a limitation and can be improved with higher resolution

1 tensorflow.org/guide/keras/rnn

parcellations and subcortical areas. Another major drawback is the network is quite ill-defined because there is no consensus what constitutes a 'cohesive' neural population. However, different atlas and network definitions seem to give similar results suggesting that the principles of BNM are at least consistent across many parcellation schemes that are used today. However, results from previous literature, show that a more detailed description of the network only improves the models performance and its ability to recapitulate rs-fMRI and that coarser models are good enough for a proof of concept application.

Moreover, another major assumption and limitation of the approach is our choice of metric, r-squared used to compare the distance between two high dimensional vectors. It assumes that better models have a higher r-squared value, although they might be explaining trivial components of the signal. Other metrics such as derivative, or the relative phase between different regions of interest might prove as a much more useful metric to compare the predictions against the empirical signal. The method also introduces another variable on when to evaluate the differences of the model. Close to the initial conditions the trajectories are too close to differentiate, and as seen from the null models where the output of the LSTM already captures a large amount of the variance in the signal. Too far from the initial conditions yields trajectories that are too far away from empirical measurements and all models become completely indistinguishable. For our results, the fourth timestep (2.88 s) was the most useful in differentiating between models, but this could vary from implementation and careful consideration needs to be used in interpreting the results and is a limitation in the approach.

4.6. Future applications

The Neural ODE techniques has a lot of potential as an additional tool in conjunction with BNM. It can be used to evaluate any differential for brain data in real time by solving for the initial conditions. Moreover, it can be used to compare across increasingly disparate brain models that are being constructed for specific applications. For individual data, it seems especially promising, since the trained network can make predictions on an individual fMRI data and thus parameters of the BNM as well as the structural connectivity can be adjusted on the individual level. Furthermore, it allows for modeling BNM trajectories during task fMRI, where the component of the signal due to network activity can be estimated, and enhance the response due to stimulus. Our results in Figure 7, indicate that the initial conditions of the Neural ODE outperform the null estimation of using the measurement as initial condition, and therefore results in trajectories that better recapitulate the short term trajectories. Therefore in the future, this algorithm can aid in separating network and task dependent activity intrinsic to fMRI.

For future approaches on more complex BNMs, it might be easier to assume the noise level and then the parameters can be solved in a more straightforward manner. The noise level seems to be endemic in the system, and rather than a parameter of the model. Since our mean surface area parcel is 858 mm^2 according to our atlas, we estimate the cortical noise per area to be $N\left(\mu = 0, \sigma = 0.35 * Surface\ Area / 858\ mm^2\right)$ and not to be a function of BNM. Once the noise level has been established, the structural perturbations are not necessary and the coefficients of the BNM can be determined directly by comparing the r-squared of the models with the empirical signal. In this manner many more complex BNMs can be compared against each other.

5. Conclusion

This manuscript investigated whether by solving for the initial conditions of a Brain Network Model for a given observation of rs-fMRI data using Neural ODE, the estimated BNM trajectories based on these initial conditions would serve as a metric to differentiate between BNMs and the measured rs-fMRI timeseries. The approach used several different FRM to fit to the rs-fMRI data by varying the global coupling, noise, and structural connectivity. The results show that the parameterization of global coupling and noise that maximizes the model's sensitivity to the structural connectivity, yields a model comparable to earlier parameterizations of the FRM. Therefore, the Neural ODE tool has the potential to differentiate and develop more complex BNMs to fit rs-fMRI data and a path to train the parameters on individual fMRI data.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Human Connectome Project.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AK: code, design, and manuscript draft. SP, PR, and SK: design, editing, and advising. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by H2020 Research and Innovation Action Grant Human Brain Project SGA2 785907 (PR), H2020 Research and Innovation Action Grant Human Brain Project SGA3 945539 (PR), H2020 Research and Innovation Action Grant Interactive Computing E-Infrastructure for the Human Brain Project ICEI 800858 (PR), H2020 Research and Innovation Action Grant EOSC VirtualBrainCloud 826421 (PR), H2020 Research and Innovation Action Grant AISN 101057655 (PR), H2020 Research Infrastructures Grant EBRAINS-PREP 101079717 (PR), H2020 European Innovation Council PHRASE 101058240 (PR), H2020 Research Infrastructures Grant EBRAIN-Health 101058516 (PR), H2020 European Research Council Grant ERC BrainModes 683049 (PR), JPND ERA PerMed

PatternCog 2522FSB904 (PR), Berlin Institute of Health & Foundation Charité (PR), Johanna Quandt Excellence Initiative (PR), German Research Foundation SFB 1436 (project ID 425899996) (PR), German Research Foundation SFB 1315 (project ID 327654276) (PR), German Research Foundation SFB 936 (project ID 178316478) (PR), German Research Foundation SFB-TRR 295 (project ID 424778381) (PR), and German Research Foundation SPP Computational Connectomics RI 2073/6-1, RI 2073/10-2, and RI 2073/9-1 (PR).

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1159914/full#supplementary-material

References

Abbas, A., Belloy, M., Kashyap, A., Billings, J., Nezafati, M., Schumacher, E. H., et al. (2020). Quasi-periodic patterns contribute to functional connectivity in the brain. *Neuroimage* 207:116387. doi: 10.1016/j.neuroimage.2019.116387

Breakspear, M. (2017). Dynamic models of large-scale brain activity. *Nat. Neurosci.* 20, 340–352. doi: 10.1038/nn.4497

Cabral, J., Hugues, E., Kringelbach, M., and Deco, G. (2012). Modeling the outcome of structural disconnection on resting-state functional connectivity. *Neuroimage* 62, 1342–1353. doi: 10.1016/j.neuroimage.2012.06.007

Cabral, J., Hugues, E., Sporns, O., and Deco, G. (2011). Role of local network oscillations in resting-state functional connectivity. *Neuroimage* 57, 130–139. doi: 10.1016/j.neuroimage.2011.04.010

Cabral, J., Kringelbach, M., and Deco, G. (2017). Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome: models and mechanisms. *Neuroimage* 160, 84–96. doi: 10.1016/j.neuroimage.2017.03.045

Chen, R., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2019). Neural ordinary differential equations. *arXiv:1806.07366v5*

Deco, G., Jirsa, V., McIntosh, R., Sporns, O., and Kötter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *Proc. Natl. Acad. Sci.* 106, 10302–10307. doi: 10.1073/pnas.0901831106

Desikan, R., Segonne, F., Fischl, B., Quinn, B., Dickerson, B., 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

Graves, A., and Schmidhuber, J. (2008). Offline handwriting recognition with multidimensional recurrent neural networks. *Adv. Neural Inf. Proces. Syst.*:545552

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159

Hjelm, D., Damaraju, E., Cho, K., Laufs, H., Plis, S., and Calhoun, V. (2018). Spatio-temporal dynamics of intrinsic networks in functional magnetic imaging data using recurrent neural networks. *Front. Neurosci.* 12:600. doi: 10.3389/fnins.2018.00600

Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci.* 104, 10240–10245. doi: 10.1073/pnas.0701519104

Kashyap, A., and Keilholz, S. (2019). Dynamic properties of simulated brain network models and empirical resting-state data. *Netw Neurosci* 3, 405–426. doi: 10.1162/netn_a_00070

Kashyap, A., and Keilholz, S. (2020). Brain network constraints and recurrent neural networks reproduce unique trajectories and state transitions seen over the span of minutes in resting-state fMRI. *Netw Neurosci.* 4, 448–466. doi: 10.1162/netn_a_00129

Mars, S. (2020). Sequence to Sequence. AI Tech & Paper Space. Avaliblie at: https://marssu.coderbridge.io/2020/11/21/sequence-to-sequence-model/ (Accessed December 03, 2020).

Mesbahi, M., and Egerstedt, M. (2010). *Graph theoretic methods in multiagent networks* (Vol. *1*). Princeton Series in Applied Mathematics; Princeton, NJ: Princeton University Press.

Nozari, E., Bertolero, M., Stiso, J., Caciagli, L., Cornblath, E., He, X., et al. (2020). Is the brain macroscopically linear? A system identification of resting state dynamics. *bioRxiv*. doi: 10.1101/2020.12.21.423856

Ritter, P., Schirner, M., McIntosh, A., and Jirsa, V. (2013). The virtual brain integrates computational modeling and multimodal neuroimaging. *Brain Connect.* 3, 121–145. doi: 10.1089/brain.2012.0120

Saenger, V. M., Kahan, J., Foltynie, T., Friston, K., Aziz, T. Z., Green, A. L., et al. (2017). Uncovering the underlying mechanisms and whole-brain dynamics of deep brain stimulation for Parkinson's disease. *Sci. Rep.* 7:9882. doi: 10.1038/s41598-017-10003-y

Salimi-Khorshidi, G., Douaud, G., Beckmann, C. F., Glasser, M. F., Griffanti, L., and Smith, S. M. (2014). Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* 90, 449–468. doi: 10.1016/j.neuroimage.2013.11.046

Sanz-Leon, P., Knock, S., Spiegler, A., and Jirsa, V. (2015). Mathematical framework for large-scale brain network modeling in the virtual brain. *Neuroimage* 111, 385–430. doi: 10.1016/j.neuroimage.2015.01.002

Schirner, M. R., McIntosh, V., Jirsa, G., and Deco, P. R. (2018). Inferring multi-scale neural mechanisms with brain network modelling. *Elife* 7:e28927. doi: 10.7554/eLife.28927

Singh, M. F., Braver, T. S., Cole, M. W., and Ching, S. (2021). Estimation and validation of individualized dynamic brain models with resting state fMRI. *Neuroimage* 226:117574. doi: 10.1016/j.neuroimage.2020.117574

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci.* 106, 13040–13045. doi: 10.1073/pnas.0905267106

Thompson, G. J., Pan, W.-J., Magnuson, M. E., Jaeger, D., and Keilholz, S. D. (2014). Quasi-periodic patterns (QPP): large-scale dynamics in resting state fMRI that correlate with local infraslow electrical activity. *Neuroimage* 84, 1018–1031. doi: 10.1016/j.neuroimage.2013.09.011

Van Essen, D., Smith, S., Barch, D., Behrens, T., Yacoub, E., and Ugurbil, K. (2013). The WU-MINN human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Vidaurre, D., Abeysuriya, R., Becker, R., Quinn, A., Alfaro-Almagro, F., Smith, S., et al. (2018). Discovering dynamic brain networks from big data in rest and task. *Neuroimage* 180B, 646–656. doi: 10.1016/j.neuroimage.2017.06.077

Wun, Z. (2020). Temporal dynamic model for resting state fMRI data: a neural ordinary differential equation approach. *arXiv preprint arXiv*

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., and Breakspear, M. (2014). Time-resolved resting-state brain networks. *Proc Natl Acad Sci U S A* 111, 10341–10346. doi: 10.1073/pnas.1400181111

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 31 August 2023
DOI 10.3389/fnins.2023.1162096

OPEN ACCESS

EDITED BY
Dajiang Zhu,
University of Texas at Arlington, United States

REVIEWED BY
Shangbin Chen,
Huazhong University of Science and Technology, China
Alexander Ho,
University of Illinois at Urbana-Champaign, United States

*CORRESPONDENCE
Yi Gao
gaoyi@szu.edu.cn

RECEIVED 09 February 2023
ACCEPTED 26 July 2023
PUBLISHED 31 August 2023

CITATION
Yang Q, Cai S, Chen G, Yu X, Cattell RF, Raviv TR, Huang C, Zhang N and Gao Y (2023) Fine scale hippocampus morphology variation cross 552 healthy subjects from age 20 to 80. *Front. Neurosci.* 17:1162096. doi: 10.3389/fnins.2023.1162096

Fine scale hippocampus morphology variation cross 552 healthy subjects from age 20 to 80

Qinzhu Yang[1], Shuxiu Cai[1], Guojing Chen[1], Xiaxia Yu[1], Renee F. Cattell[2,3], Tammy Riklin Raviv[4], Chuan Huang[5,6], Nu Zhang[7] and Yi Gao[1,8,9]*

[1]School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, Guangdong, China, [2]Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY, United States, [3]Department of Radiation Oncology, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY, United States, [4]The School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er Sheva, Israel, [5]Department of Psychiatry, Stony Brook University, Stony Brook, NY, United States, [6]Department of Radiology, Stony Brook University, Stony Brook, NY, United States, [7]Department of Neurosurgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, Guangdong, China, [8]Shenzhen Key Laboratory of Precision Medicine for Hematological Malignancies, Shenzhen, China, [9]Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China

The cerebral cortex varies over the course of a person's life span: at birth, the surface is smooth, before becoming more bumpy (deeper sulci and thicker gyri) in middle age, and thinner in senior years. In this work, a similar phenomenon was observed on the hippocampus. It was previously believed the fine-scale morphology of the hippocampus could only be extracted only with high field scanners (7T, 9.4T); however, recent studies show that regular 3T MR scanners can be sufficient for this purpose. This finding opens the door for the study of fine hippocampal morphometry for a large amount of clinical data. In particular, a characteristic bumpy and subtle feature on the inferior aspect of the hippocampus, which we refer to as hippocampal dentation, presents a dramatic degree of variability between individuals from very smooth to highly dentated. In this report, we propose a combined method joining deep learning and sub-pixel level set evolution to efficiently obtain fine-scale hippocampal segmentation on 552 healthy subjects. Through non-linear dentation extraction and fitting, we reveal that the bumpiness of the inferior surface of the human hippocampus has a clear temporal trend. It is bumpiest between 40 and 50 years old. This observation should be aligned with neurodevelopmental and aging stages.

KEYWORDS

hippocampus, fine-scale segmentation, shape analysis, deep learning, MRI

1. Introduction

Numerous radiological studies of sub-cortical morphology have shown many brain disorders to be correlated with hippocampal shape (Styner et al., 2004; Thompson et al., 2004; Apostolova et al., 2006; Wang et al., 2006; Scher et al., 2007; Colliot et al., 2008; Nestor et al., 2013; Gao et al., 2014; Gao and Bouix, 2016), volume (Fleisher et al., 2008), or metabolic properties (Kraguljac et al., 2013). The hippocampus also exhibits important related variations in healthy individuals. For example, spatial memory declines with age and this is consistent with a decreasing trend in hippocampal volume (Bohbot et al., 2004; Konishi et al., 2017). Moreover, the hippocampal structure also correlates with the function of establishing semantic associations in memory (Henke et al., 1999). As people age, the

rate of hippocampal atrophy increases, with the greatest increase after middle age (Fraser et al., 2015). These comparable global features of non-clinical and clinical conditions (Convit et al., 1997; Schuff et al., 2009) provide an important measurement for the evaluation of hippocampal abnormalities and functions.

Morphological and functional assessment of fine-scale structures are still considered challenging tasks. The hippocampus is known to be among the few structures where neurogenesis continues to take place after birth. Similar to the formation of any other cortical gyrus/sulcus, the proliferation and stacking of cells in hippocampal neuronal layers requires space-efficient outward folding of the hippocampal surface. Furthermore, it is worth noting that hippocampus neurogenesis-associated features exhibit both qualitative and quantitative age-related alterations (Knoth et al., 2010). This work aims to investigate the macroscopic morphological appearance and its age-dependent variability across the life span of the hippocampus.

The structure of hippocampal dentation is of particular interest due to its apparent rugged ridges, which are in the CA1/subiculum on the inferior aspect of the hippocampal body and extend through the inferior medial aspect of the tail (Duvernoy, 2013). Consulting neuroanatomy textbook (Duvernoy et al., 2005; Arslan, 2014; Ribas, 2018; ten Donkelaar et al., 2018), the dentated appearance is obvious and exhibits great variability of shape as shown in Figure 1. Unfortunately, this variability has been largely overlooked in previous image based studies.

Such morphological variation mostly involves the CA1 regions. CA1 neurons are known to be involved in episodic memory (Bartsch et al., 2011) and a positive correlation between cortical gyrification and cognitive functioning was found (Luders et al., 2008). Further quantitative studies related to episodic memory (Beattie et al., 2017) have used ultra-high resolution MRI data to explore the highly variable long axis of hippocampal dentation and its functional role in episodic memory.

Quantitative feature generation would be a valuable tool for the intuitive, concise, and personalized characterization of hippocampal dentation. Moreover, hippocampal dentation varies across individuals, over time and along the inferior surface. This variation makes it significant for quantifying the relationship between hippocampal dentation and other factors, such as age, clinical, or non-clinical conditions. More importantly, the quantitative analysis of fine-scale structures allows us to leverage advanced machine learning methods and enables us to explore data sets more extensively.

There are two main reasons why current research is insufficient to quantify hippocampal dentation changes. First, quantitative research methods usually require a large data size, but the limited acquisition of high resolution image data with hippocampal fine-scale structure leads to difficulties in large-scale research. The main reason for this is because clinical 3T scanners find it difficult to acquire sufficient resolution and the currently finite availability of ultra-high field scanners (7T or greater) (Wisse et al., 2012; Kim et al., 2013; Derix et al., 2014) or post-mortem specimens (Yushkevich et al., 2009). Second, compared to global structure, fine-scale structure is difficult to characterize by most handcrafted feature representation in feature engineering (Bengio et al., 2013) or automatic extraction of features through deep learning networks. This may be due to the small, hard-to-measure structural geometry and the challenge of properly delineating regional boundaries. Additional challenges stem from dentate variability along the different sagittal slices of hippocampal dentation.

The above aspects have made it difficult to conduct a quantitative analysis of the dentated shape of the hippocampus. The most closely related work by Kilpattu Ramaniharan et al. (2022) visualized dentation after using the up-sampling method and ASHS software. They counted dentation and explored its association with memory dysfunction in patients with temporal lobe epilepsy that have hippocampal sclerosis. Beattie et al. (2017) visualized the dentation using ultra-high resolution structural MRI and using a visual rating scale, accessed by human observers, which showed that the extent of dentation varied considerably across individuals and was positively correlated with memory recall and visual memory recognition. The raters in that study needed to examine all sagittal slices to observe dentation visible through the entire width of the hippocampus. This work is labor-intensive, highly subjective, and can suffer from high intra- and inter- reader variability. Therefore, this rating scheme cannot be generalized reliably to a large number of subjects across multiple institutions. A computed aided quantification and analysis framework for evaluating the hippocampal dentation is therefore needed to provide objective fine-scale morphometry.

This analysis framework consists of two components. First, an effective and efficient segmentation algorithm is needed that is capable of capturing fine-scale dentations. It has been shown previously that such local and subtle features under the hippocampus can be reconstructed from clinical 3T MRI by a multi-atlas based technique (Chang et al., 2018). However, the multi-atlas warping technique (Nestor et al., 2013) could not fulfill the further need for a large population study due to it being extremely time-consuming. More recently, the use of a 3D deep convolutional neural network for hippocampus segmentation has achieved high precision, measured by a global metric such as the Dice coefficient (Thyreau et al., 2018). Part of its training labels was from the FreeSurfer algorithm (Fischl, 2012). Using the synthetic data and augmentation algorithm, the Dice average coefficients are above 90%. Later, in a hippocampal segmentation study of a stroke population, Zavaliangos-Petropulu et al. (2022) used deep learning based Hippodeep method (Thyreau et al., 2018) and make a comparison with FreeSurfer. Rather than achieving annotation with the help of Freesurfer, Goubran et al. (2020) trained the CNN using 259 bilateral manually delineated segmentations to achieve better performance. Guo et al. (2020) proposed a longitudinal classification-regression model for segmenting the hippocampus in infant brain MRIs. Work by Liu et al. (2020) proposed a joint automatic hippocampal segmentation and AD classification method. For refined segmentation by exploiting space information, Pang et al. (2019) proposed a method based on iterative local linear mapping (ILLM) with representative and local structure-preserved feature embedding. To improve segmentation quality, Van Opbroek et al. (2018) and Ataloglou et al. (2019) explored different transfer learning techniques.

Even though current CNN based hippocampus segmentation methods have achieved global accuracy measures, they still lack the ability of the 3T images to capture fine-scale dentations.

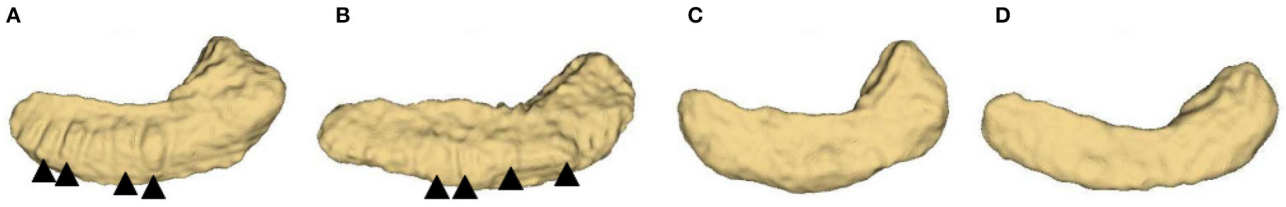


FIGURE 1
Inferior view of four hippocampus. **(A, B)**: bumpy group; **(C, D)**: smooth group. The arrowheads indicate the prominent dentations and their approximate orientations.

Moreover, to conduct large scale statistical morphological studies, the generalization capability of the CNN needs to be strengthened to handle the image intensity fluctuations among different scans, machine-dependent noise, and bias field in-homogeneity, etc. To address such issues, Memmel utilized the data from different domains with the GAN framework to disregard domain-specific information (Memmel et al., 2021). For hippocampus segmentation across different datasets, few studies have considered how to solve this with the help of domain adaptation in an end-to-end framework directly. In response, our proposed framework can be adopted in these similar studies. To further improve the framework, some research by Strudel et al. (2021) and Valanarasu et al. (2021) initially used the transformer block to extract features and process the long range relationship of these features. To utilize these techniques we need to improve the feature extraction ability of the design framework before subsequent discriminator and segmentation steps, noting that based segmentation operation is the foundation to help for the subsequent fine-scale segmentation and accurate shape analysis.

Once the hippocampus is successfully segmented and its fine-scale morphology features are extracted, we need to design a technique that specifically compares the dentated structures underneath the CA1 region. Previously, the analysis of shapes is usually conducted between two groups of shapes, trying to identify the region where the two groups of shapes differ significantly (Gerig et al., 2001; Shen and Makedon, 2006; Styner et al., 2006; Cates et al., 2008; Shen et al., 2009; Shen, 2010; Riklin Raviv et al., 2014; Hong et al., 2015; Gao and Bouix, 2016). However, the scenario is different in this work as we have already identified certain regions on the shape, as well as the possible pattern of variation. We are more explicitly interested in the magnitude of the dentated pattern between the two groups. To this end, we have to design a suitable approach to handle the problem at hand.

As an exploratory proposition, we hypothesize that the level of dentations may be involved in neurogenesis with age, reflected by variation of dentated structure along its long axis. This work presents a novel domain adaption segmentation and regression model of quantitative features on a relatively large dataset of 552 subjects (1,104 hippocampi) (IXI dataset, 2018). As a key step in the successful application of machine learning for quantitative estimation (Bengio et al., 2013), to handle the great variability of hippocampal dentation, we combined the advantages of domain adaption segmentation in the field of deep learning and propose a new feature representation method for dentation analysis.

Using deep learning methods, we have designed a transformer based approach to segment and extract the hippocampus. This approach is then combined with the learned grayscale information of the hippocampus, and multi-scale segmentation is performed to obtain fine-scale segmentation. Once the dentated structures are extracted, we then measure the magnitude of the dentation structure by first identifying the long axis of the hippocampus. This is done under a point cloud representation of the shape. Then, specifically engineered for the dentation under the CA1 region, a non-linear fitting of the sinusoidal function is performed; the observation that the dentation presents an arciform or sinusoidal appearance allows us to quantify the convolution by magnitude and frequency of the sinusoidal function. Moreover, using simulated annealing, we can find the most optimal model parameters.

Our work contributes to the field in three ways: (1) A deep-learning based robust segmentation algorithm is used to extract the fine-scale hippocampal morphological feature at the sub-pixel level on a large dataset. (2) This study demonstrates that certain fine-scale hippocampal morphological features vary with aging. (3) To our knowledge, even though rich hippocampal shape studies have been conducted previously, this is the first fine-scale quantitative analysis on hippocampal dentation based on a clinically available dataset. Our method aims to study the differences in hippocampal shape of a healthy population over an age range from people aged in their mid-20s to 80 years old. The construction of an analytical baseline and the development of a technique for robust and quantitative image analysis will open possibilities for future comparisons between non-clinical and clinical groups.

2. Materials and methods

To use the hippocampus segmentation algorithm, 41 3T MR images with the hippocampus manually traced out were used to train the segmentation model. These 41 cases are from the EADC-ADNI Harmonized Protocol project (Apostolova et al., 2015; Boccardi et al., 2015; Frisoni et al., 2015). We also performed segmentation validation with real 7T MR images based on samples from Alkemade et al. (2020). For aging hippocampus morphology research, the hippocampi of 552 healthy subjects (age range: 20–79, mean age = 48.2 ± 16.0 years) from the IXI dataset were analyzed (IXI dataset, 2018). The age distribution of the subjects is summarized in Table 1. The 1,104 3T T1 weighted MP-RAGE MR images of 552 subjects were used. Before data analysis, all the scans

were resampled to isotropic 1 mm resolution, and we named that native resolution.

As shown in the flowchart in Figure 2, the main pipeline consists of three parts. First, an automatic and robust segmentation algorithm was proposed to capture the fine-scale morphology of the hippocampus based on common 3T MR images. Then, the characteristic fine-scale dentation is extracted from the segmentation through a non-linear regressor. After that, the level of the dentations is quantitatively analyzed against age to identify temporal changes. In what follows, we detail all three major algorithm components.

2.1. Fine-scale semantic segmentation

The multi-atlas-based methods used in Chang et al. (2018) have the disadvantage of involving a large computation time when applied to a large-scale dataset for study. Therefore, we propose a new deep-learning-based fine-scale segmentation method to obtain fine-scale dentation of the hippocampus from 3T MR scans relating to 552 subjects. However, there are some serious issues to solve before such fine-morphological analysis can be efficiently applied to such a large cohort.

TABLE 1 The amount of MRI acquisition subjects in each age range from the IXI dataset.

Age range (years)	Num. subjects
20–29	100
30–39	99
40–49	89
50–59	98
60–69	117
70–79	49
Total	552

The first issue is that the large amount of image data used to segment the multi-atlas-based approach used in previous approaches, such as Chang et al. (2018), is too time-consuming to be practically useful. Toward this goal, the recent development of deep learning methods provides a promising alternative to multi-atlas approaches.

Second, the large IXI cohort analyzed does not have any manual annotation. We, therefore, need to utilize carefully validated annotation from the EADC-ADNI dataset (Apostolova et al., 2015; Boccardi et al., 2015; Frisoni et al., 2015) for training, and apply the trained model to the IXI dataset. This inevitably introduces a cross-dataset discrepancy between the training and testing images, and adequate domain adaptation is necessary.

Third, and most importantly, even the expert-curated hippocampus annotation in Apostolova et al. (2015), Boccardi et al. (2015), and Frisoni et al. (2015) does not capture the fine-scale hippocampus dentations and a deep learning model trained on such annotation is not capable. To perform the sub-pixel fine-scale morphometry, we have to depart from the constraint of the learned space and extend the segmentation to a much higher resolution level.

To address the above issues, in this sub-section, we propose the hippocampal domain adaption fine-scale segmentation method to capture the fine-scale hippocampus dentation structure from the clinically available 3T MR images. The algorithm pipeline is illustrated in Figure 3.

2.1.1. Semantic segmentation

The deep-learning-based semantic segmentation method formulates the dentation annotation task as a pixel-classification problem. The core encoder-decoder framework consists of a stack of sequentially connected convolutional layers and long-range skip connections. The locations of the feature information in a

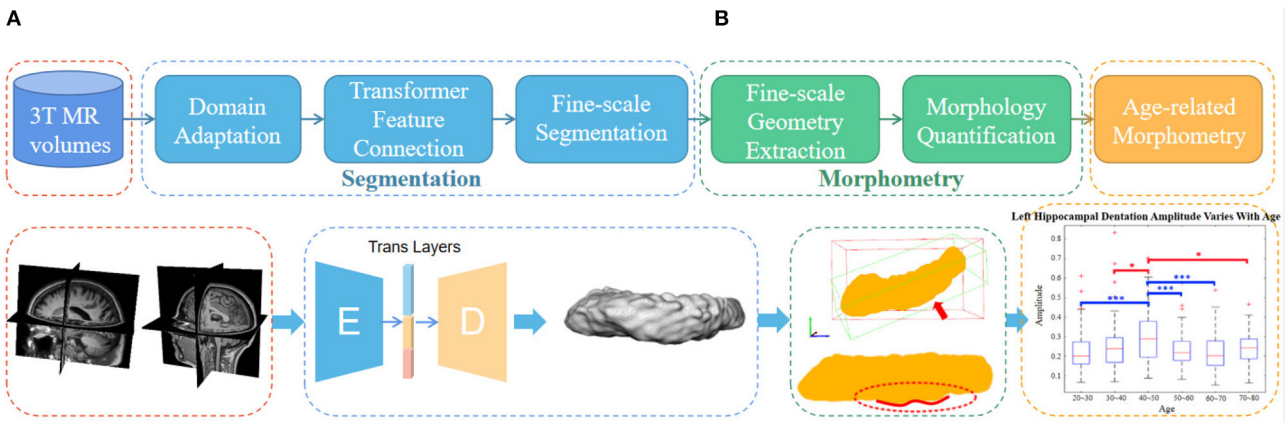


FIGURE 2
The overall flowchart of the proposed fine-scale segmentation and morphometry. **(A)** Hippocampus are segmented from the 3T MR images and the dentation and obtained from fine-scale segmentation. **(B)** After that, geometry features are extracted and age-based associations are explored. Dotted boxes of the same color indicate that they belong to the same content in the pipeline. Note that this only represents the overall flow of the manuscript. The detailed and high resolution figures for each portion here will be shown in the subsequent part.

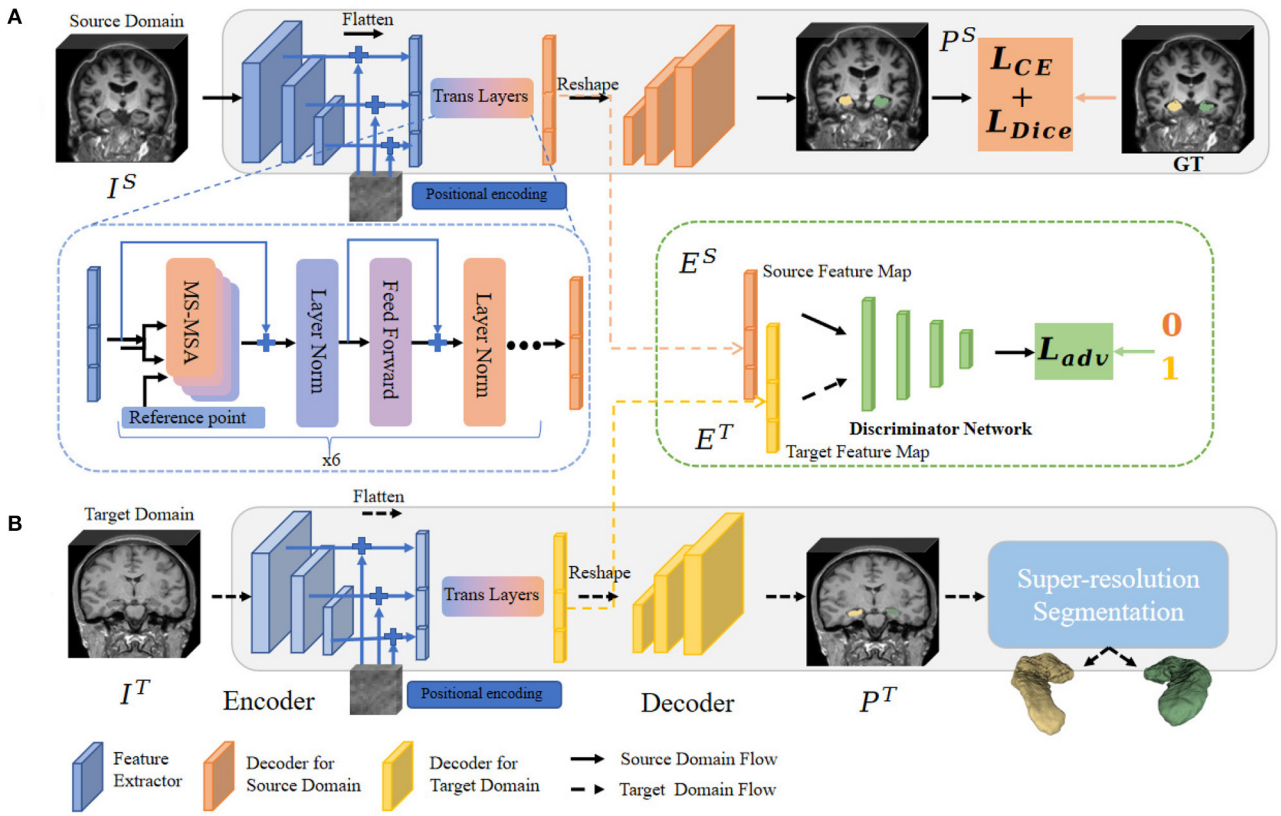


FIGURE 3
Overview of the fine-scale segmentation framework. The 3D MR images in the source and target domain pass through procedures **(A, B)**, respectively. In procedure **(A)**, CE and Dice Loss are computed based on source annotation to optimize the semantic segmentation framework. The multi-feature from different level layers also passes through the MSA layers and outputs the convergence feature. Next, images from the target domain were fed into procedure **(B)**. Before obtaining a fine-scale segmentation map, we adopt a module to constrain the extracted feature map between two different domains. The domain adaptation operation is illustrated in the block with a green edge. Finally, the possibility map from the target domain is forwarded through fine-scale segmentation (the blue block) to output the final annotation.

higher layer are computed based on the locations of tensors of the next lower layers as they are connected through a layer-by-layer up-sampling operation. However, due to the locality nature of the convolution operation, the receptive field is limited along with the depth of layers and the size of the convolutional kernel. As a result, only higher layers with big receptive fields can model long-range dependencies in the vanilla encoder-decoder architecture. More recently, the multi-head self-attention mechanism (MSA) of the vision transformer shows a more effective strategy for learning long-range contextual information. As a result, we utilize a transformer-based MSA framework to overcome this limitation, which is motivated by Xie et al. (2021).

As shown in Figure 3, we bridge the transformer layer to the design of encoder architecture and aim to help engage lower and higher contextual features directly and capture the long-range dependency of pixels effectively. With such an encoder partition, the multi-scale features extracted from convolution are concatenated before being forwarded through MSA. However, Xie et al. (2021) sets the hidden size in residual blocks of the hierarchical encoder to 384 to keep the same hidden size in the feed forward network of MSA. At the same time, the small kernel size 3 used shows a lower capture ability, while the larger kernel size can capture dependencies between information units further away in the earlier layers. It only accepts inputs of the same size ($48\times192\times192$), which is not conducive to the segmentation of small organs, such as the hippocampus. In contrast to Xie et al. (2021), we squeeze the channels of the residual blocks to 192 as half of the original 384 channels and further adopt several groups of larger kernel-sized convolutions to expand the dependencies capture ability of inner-place units. Moreover, we engage the last three layers of contextual feature output from the encoder together to get finer-scale spatial information. In short, we improve by replacing the channel complexity with spatial complexity. We also modified the size limit of the input to be able to take a smaller size of 64^3 than $48\times192\times192$ in dimensions 2 and 3 as input and focus more on the target hippocampal region. In the next step, these extracted features are passed to the MSA layers to aggregate hierarchical long-range dependency.

It should be mentioned that there is a mismatch between the 3D image tensor and the 1D sequence when bridging the transformer layer. As linear projection processes the information in a sequence-to-sequence manner, the feature maps produced by the encoder from every stage must be flattened into a 1D sequence before feeding into transformer layers. Also, it has to face the problem

of losing spatial information when it is being flattened. So we add the 3D positional encoding sequence to supplement position information to solve this problem. Furthermore, to improve the computational efficiency, we utilize the set of key sampling locations (denoted as r_p) in the image around the reference location. As a result, the MSA layers can be formulated as:

$$MSA\left(\{f_l\}_{l=1}^{L}, z_q, r_p\right) = \Psi\left(Concat\left(h_1, h_2, \ldots, h_N\right)\right) \quad (1)$$

where N is the number of the heads (denoted as h_i and set as 6), and flattened feature maps $\{f_l\}_{l=1}^{L}$ are extracted from the L stages of the left encoder. z_q is the feature representation of query q, which is gotten from $\{f_l\}_{l=1}^{L}$ and position embedding feature, $\Psi(\cdot)$ is the Linear projection operation to weight and aggregate the features.

In the Decoder part, the output sequences of transformer layers are separated and reshaped according to the size of feature maps from the encoder at each stage. The processed features from each stage are then concatenated with processed features from the deconvolutional layers of the preceding stage. Finally, they are fed into a residual block followed by a $1 \times 1 \times 1$ convolutional layer with a proper activation function (softmax) for computing the segmentation probabilities of the hippocampus.

To efficiently illustrate the workflow, we denoted the semantic segmentation framework as S. It takes the cropped sub-volumes of image volumes from Source Domain (denoted as $I^s : \Omega^3 \subset \mathbb{R}^3$) as input, and generates an output of the same shape (denoted as $P^S : \Omega^3 \subset \mathbb{R}^3$). Plus, all these volumes were resampled to isotropic of 1 mm^3 before segmentation. The corresponding annotation of I_i^s are also denoted as $L^s : \Omega^3 \subset \mathbb{R}^3, \Omega^3 \to 0, 1, 2$. In order to optimize S and get better parameter W_S, we utilize the segmentation loss $\mathcal{L}_s$, which is defined as:

$$\mathcal{L}_s = \lambda_1 * \underbrace{\frac{1}{N}\sum_{i=1}^{N} -y_{ij}\log\left(p_{ij}\right)}_{\mathcal{L}_{ce}} + \underbrace{\frac{1}{N}\sum_{c=1}^{N} -\frac{2\sum_i p_{ij}y_{ij}}{\sum_i p_{ij} + \sum_i y_{ij}}}_{\mathcal{L}_{dice}} \quad (2)$$

where p_{ij} and y_{ij} refer to the segmentation predicted probability and corresponding category segmentation for voxel i, j. N means the voxel number. The segmentation loss function can be minimized end-to-end by getting optimized W_S. Finally, the output channel of the network is set as 3, for the left, the right hippocampus, and the background. This semantic segmentation framework is intended for images from the ADNI dataset with observed distribution, and the next step is to address the problem of obtaining hippocampal annotation for images from the IXI dataset with unobserved distribution.

2.1.2. Image normalization through domain adaption

The MR images of the large IXI cohort to be analyzed are acquired from different machines and are of different protocols from the training MRI cohort where the hippocampus is labeled. Since MR images across machines do not share reference voxel values, the training images may have different intensity values and/or texture patterns from the testing ones.

As a result, we have to normalize the distribution of the images from the IXI dataset (the target domain) with those in the training and validation ADNI datasets (the source domain). To that end, we train our segmentation model with a discriminator network to make the adaption between the two sets.

Denote the two sets of images from the source and target domains as I^s and I^t, respectively. We forwarded the source image I^s to the semantic segmentation network S and calculate the difference between output and annotations for an optimal S. Then, we predicted the segmentation output P^t for the target image I^t. Since our goal is to make segmentation predictions P^s of source and P^t of target images close to each other, we used these two predictions from the segmentation framework as the input to the discriminator D to distinguish whether the input is from the source or target domain.

Optimizing the adversarial loss on the target prediction, the network propagates loss gradients from D to S, which encourages S to generate similar segmentation distributions in the target domain to the source prediction. With the proposed method, we formulate the adaptation task containing discriminator loss functions:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{adv}\mathcal{L}_{adv} \quad (3)$$

where $\mathcal{L}_{seg}$ is the semantic segmentation loss using ground truth annotations in the source domain, and $\mathcal{L}_{adv}$ is the adversarial loss that adapts predicted segmentations of target images to the distribution of source predictions. The λ_{adv} is the weight used to balance the two losses.

For the discriminator, we use an architecture similar to Tsai et al. (2018) but utilize the extracted feature from transformer layers and the final softmax segmentation possibility map to explore more spatial information. Furthermore, only one discriminator is utilized in our framework. The discriminator network consists of convolution layers followed by an adaptive average pooling unit and a full connection layer for the binary classification as illustrated in Figure 3 (green block). The discrimination cross-entropy loss can be written as:

$$\mathcal{L}_{adv}(E) = -\sum\left((1-z)\log\left(D(E)\right) + z\log\left(D(E)\right)\right) \quad (4)$$

For discrimination, z is set to 0 if the sample is gotten from the source domain and z is set to 1 if the sample is from the target domain.

2.1.3. Fine-scale segmentation

Although a rich amount of deep-learning based hippocampus segmentation schemes exist, one of their shortcomings is that at the native image resolution the fine dentation morphology is not captured in the manual annotation. Moreover, as the deep-learning based methods depend more on the training annotation, it is apparent that if certain shape features do not exist in the training set, it is unlikely they will be captured accurately in the external testing images.

On the other hand, the probability map of the hippocampus obtained above contains valuable information about the approximate morphology. Since we aim to extract the fine morphology, it is valuable to escape the realm learned by the

deep learning framework and explore the fine-scale morphology features unseen in the training images. This is detailed below.

To refine the surface of the hippocampus, we employ the fine-scale method (denoted as $SR(\cdot)$) to fine-tune the probability map $M^T : \Omega^s \rightarrow [0, 1]$ from S at the native resolution. However, when such a data-driven method is carried out in a high-resolution space, it will consume more than 100 times in computer memory and complexity compared to native resolution. To reduce such a burden and make the computation practical, we only handle the region of interest focal to the hippocampus according to the possibility map M^T of target domain image I^T to get R^T. This step of cropping is performed automatically by the program. By doing this, we can perform segmentation on a sub-millimeter level morphological feature contained in the grayscale information and get the final result S^{sr}. The operation can be denoted as:

$$S^{sr} = SR\left(Crop\left(M^T\right)\right) \tag{5}$$

More explicitly, we get the sampled observation in isotropic 0.2 mm/voxel resolution by the factor H being set as 5 and up-sampled the images through convex 3D interpolation while balancing the consumption and efficiency in the morphology study. To construct the hippocampus in high resolution space with fine location, first, we define the high confidence region as $C \coloneqq \left\{x \in \Omega^{s/H} : M^T(x) > \eta\right\}$, where higher $\eta \in [0, 1]$ values indicates the voxel belonging to hippocampus with higher confidence and $\Omega^{s/H}$ is the new images after H times cubic spline up-sampling. However, such a strongly constrained C does not make full use of the hippocampal surface context information of images in the high image space, which might even crudely omit some dentations.

To address this issue, we used the following variational approaches to refine the hippocampal surfaces. We denote the family of evolving surface as $\zeta \subset \mathbb{R}^3$, $\zeta = \partial C$, and for surface optimization, we define an energy functional as:

$$\mathbb{E}(\zeta) \coloneqq -\int_{x \text{ in } \zeta} \alpha M^T(x)dx + \beta \int_{\zeta} dA \tag{6}$$

where the x traverses the space inside the closed surface ζ, and the joined $\int_{\zeta} dA$ is the total surface area. The α and β are the positive weight. Calculating regional statistic force and edge-based force, the flow of the surface is controlled by the partial differential equation below:

$$\frac{\partial \zeta(\boldsymbol{p}, t)}{\partial t} = [L((\boldsymbol{p}, t)) - \alpha M^T(\zeta(\boldsymbol{p}, t)) + \beta \nu(\boldsymbol{p}, t)]\mathbf{V}(\boldsymbol{p}, t) \tag{7}$$

where $\mathbf{V}$ is defined as the inward unit normal vector field on ζ, $\boldsymbol{p}$ is the spatial parameterization of surface and ν is the mean curvature of the surface. In Equation (7), Laplacian of Gaussian function (LoG) is defined as $L((\boldsymbol{p}, t))$ for edge based force. To balance the force of edge evolution, the joined term on the right of LoG is the regional statistic force. The surface optimizing Equation (7) does not necessarily reside in the learned space of the neural network. This means it escapes from the learned space, which does not have the fine morphology, and the surface evolves and converges to the locations that process strong edge appearance and close to the probability map M^T with high confidence (control by setting η value).

Both the deep-learning and fine-tuning processes above are fully automated. As a result, the final surface will not only achieve high local similarity measures such as the Dice coefficient but will also successfully capture the fine-scale hippocampal dentations, which is the critical shape feature for subsequent morphology studies.

2.2. Dentation feature extraction and analysis

The goal of dentation analysis is to quantitatively explore the denotational shape variation between different groups. To this end, we first extract the dentation region by projecting the shape to a proper plane. Then, the dentation could be modeled as sinusoidal curves, whose parameters are obtained by non-linear fitting. Once the parameters of the curves have been found, the dentations across different groups are compared.

As can be seen from Figure 4, the dentations reside on the inferior surface of the CA1 section and are one or more relatively parallel ridges. Based on such observation, if one could project the 3D dentation structure along its ridge, the resulting 2D silhouette should have a sinusoidal appearance. It is workable to only capture the magnitude and frequency of such sinusoidal waves to characterize the dentations. Following the ideas above, the proposed method contains the following steps.

2.2.1. Point cloud representation

The fine-scale segmentation method provides a very detailed extraction of the hippocampal structure and allows a detailed analysis of the dentation structure. Following the ideas above, we first project the 3D shape to a plane that optimally reveals the dentation in the 2D plane. To aid the projection step, we represent the extracted hippocampus using a point cloud. Following Gao and Tannenbaum (2010), the point cloud is a collection of data points defined by a given coordinate system, which carries the morphological information of hippocampal structure. The segmented hippocampus region can be denoted as a binary image $J : \mathbb{R}^3 \rightarrow 0, 1$. J can therefore be considered as a probability density function (pdf) of a random variable which uniformly distributed in the hippocampus region. Next, we extract samples from such a pdf. Due to the irregular shape of J's support, we employ the rejection sampling for the sample extraction. As a result, each hippocampus is represented as a cloud of points $X = x_i \in \mathbb{R}^3$ that are further processed in the subsequent sections.

2.2.2. Medial axis representation

Looking laterally, the dentation features (or lack thereof) are evident underneath the gyrus region. One needs to project the 3D shape to the correct 2D view, identify the inferior boundary, and then quantitatively represent the dentation feature for comparison across ages. In this context, principal component analysis (PCA) is a suitable and effective linear dimension reduction technique to serve the purpose of extracting the AP axis and the inferior surface of the hippocampus.

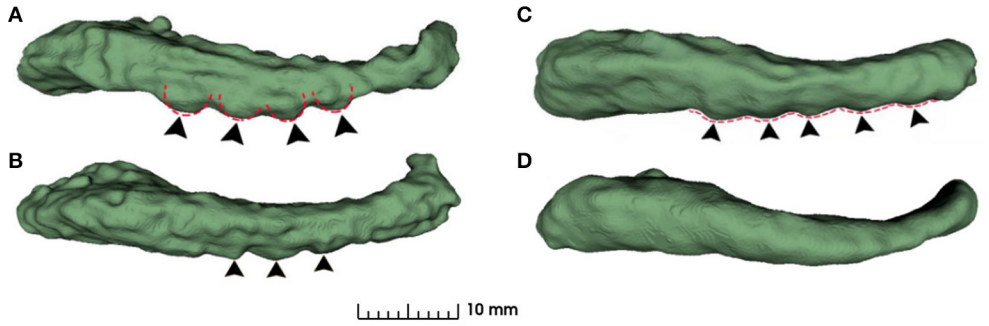


FIGURE 4
Examples of variation in degree of dentation. **(A)** Depicts a high degree of dentation with many prominent, or arciform dentations. **(B)** Depicts a few arciform dentations. **(C)** Shows few, less prominent sinusoidal dentation. **(D)** Shows a hippocampus with no apparent dentation. Arrowheads indicate individual dentation. Dashed red lines illustrate the contour of arciform **(A)** and sinusoidal **(C)** dentation.

In practice, PCA projects the data points to the subspace, which maximizes the variance to keep the structure of interest in the volume as much as possible. The first few eigenvectors that yield the largest eigenvalues often explain most of the variance in the data. The shape of the hippocampus has the longest axis in the AP direction. After that, the lateral width of the hippocampus is about 3 cm whereas the thickness in the superior-inferior direction is the smallest, which is less than 1 cm. As a result, when performing the PCA on the points in the hippocampus, the eigenvector corresponding to the largest eigenvalue is expected to be roughly the AP direction but slightly tilted up. Following that, the second mode should be in the left-right direction and the third eigenvector should be perpendicular to the "sheet" of the hippocampus.

As a result, if we project all the 3D points in the hippocampus along the second direction onto the plane spanned by the first and third eigenvectors, we could observe the dentations clearly in the 2D view. This is shown in Figure 5.

2.2.3. Sinusoidal dentation modeling and fitting

As shown in Figure 5, hippocampus dentation on its inferior surface is observed to have many arciform or sinusoidal prominence in the dotted red circle. A similar variation pattern in 3D is reflected in the 2D geometry after the projection. This variation of appearance can be approximated by a sinusoidal fitting model, which allows for a quantitative description by measuring dentation with two parameters of magnitude and frequency. The core of the quantitative analysis is to find model parameters that can reflect the prominence of dentation in the hippocampal sub-region, which has inter-subject variation. In this work, a non-linear fitting model was established to explore and characterize this simple dentation variation. We compute the model parameters that lead to an optimal adaptation of the variation to the set of observations. Specifically, we fit a sinusoidal function to the silhouette of the inferior surface of the hippocampus and measure the parameters of the sinusoidal function. Mathematically, we build a two-dimensional Cartesian coordinate for each hippocampus. This coordinate has its origin at the center of mass of the hippocampus. Next, two axes are pointing to the eigenvectors with the largest and smallest eigenvalues from the PCA method, respectively. Visually, this forms a plane cutting through the hippocampus vertically along its major axis. All points in the hippocampus are projected onto this plane, forming a 2D region as shown in row 3 of Figure 5.

The silhouette of the inferior hippocampal surface is therefore denoted as a function in this coordinate system. The sinusoidal fitting is cast as an optimization problem:

$$J(\boldsymbol{A}, \boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{b}) := \int_x \left(y(x) - [\boldsymbol{A}\sin(\boldsymbol{w}x + \boldsymbol{\phi}) + \boldsymbol{b}]^2\right) dx \quad (8)$$

In Equation (8), there are four fitting parameters: amplitude ($\boldsymbol{A}$), frequency ($\boldsymbol{f} = \boldsymbol{w}/2\pi$), phase ($\boldsymbol{\phi}$), and bias ($\boldsymbol{b}$). Among them, amplitude and frequency are the two key parameters that describe the height and the density of dentations.

For ease of understanding, we will obtain the height and width of the hippocampal dentation and display it graphically. As shown in Figure 5E, the height H is twice the amplitude $\boldsymbol{A}$, so $H = 2\boldsymbol{A}$, and the hippocampal bump width can be expressed as $L = 1/\boldsymbol{f}$. Therefore, the goal is to find parameters to minimize J to obtain the optimal parameter magnitude and frequency and this is a non-linear optimization problem.

In order to address this non-linear optimization problem, simulated annealing (SA) (Khachaturyan et al., 1981) was employed to find the optimal parameters $\boldsymbol{A}$, $\boldsymbol{w}$, $\boldsymbol{\phi}$, $\boldsymbol{b}$. SA is a probabilistic approach for getting the proximate global optimum of a given non-linear function. Compared with the general greedy algorithm, the SA introduces random factors, which may accept a solution worse than the current solution with a certain probability. This means that SA is able to jump out of the local optimal solution and approximate the global optimal solution.

The following annealing criteria are used to allow for accepting a "worse" solution:

$$\mathbf{e}^{-\Delta D/T} > R(0,1) \quad (9)$$

where ΔD is the difference of cost implied by the balance, the temperature is initialized high and gradually "cool" to simulate the heating process, and $R(0,1)$ is randomly distributed on [0,1].

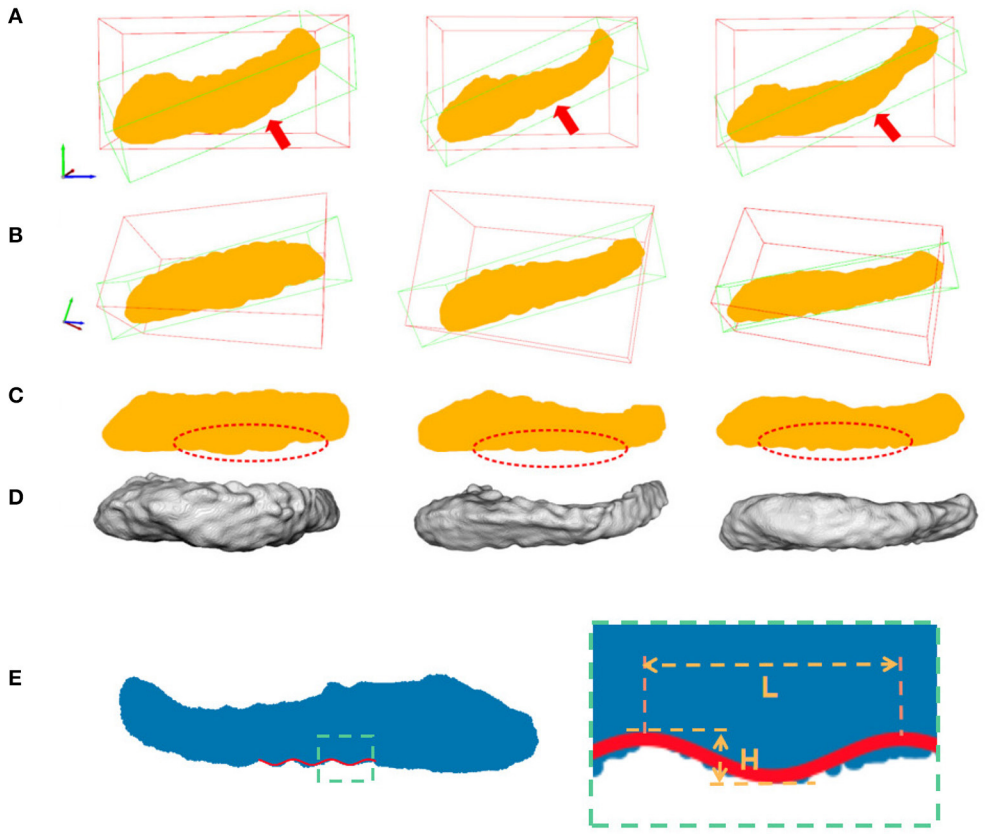


FIGURE 5
Hippocampal 3D point cloud representation, with the resulting plane of interest. Rows **(A, B)** show the hippocampal results from fine-scale segmentation. Row **(C)** shows the corresponding point cloud representation of hippocampus. The red arrows showed the second major sight of view in 3D space. The red bounding boxes are axis-aligned bounding boxes of point cloud and the green bounding boxes are oriented bounding boxes based on the PCA of their convex hull. Row **(D)** are results after dimension reduction processing and dentation can be seen in it. Row **(E)** shows measurements of the height and width of hippocampal dentition after PCA. H represents the peak and trough distance of the curve, and L represents the length of one bump of the curve.

2.3. Experiments and evaluations

2.3.1. Experimental setup

For this study, obtaining accurate fine-scale segmentation results is the prerequisite for accurate longitudinal analysis. It is therefore critical to evaluate the fine segmentation performance of the proposed framework. To that end, the proposed framework was compared against the following state-of-the-art (SOTA) techniques: Hippodeep (Thyreau et al., 2018) proposed a CNN trained on hippocampal segmentation from multiple cohorts including 2,500 T1 MR scan images. HippMapp3r (Goubran et al., 2020) proposed and trained a 3D CNN using 259 bilateral manually delineated segmentation acquired at multiple sites on different scanners. In addition, to acquire better performance of encoder and decoder, we utilized the based segmentation framework proposed in Xie et al. (2021). Therefore, it is also included in comparison experiments.

We performed our experiments on the two datasets. The EADC-ADNI dataset (Apostolova et al., 2015; Boccardi et al., 2015; Frisoni et al., 2015) containing 41 manually labeled subjects was randomly divided into a training group (N = 30) a validation group (N = 11). IXI dataset (2018) contains 552 subjects and all of them were utilized as test groups. The proposed frameworks were trained on the training group, the validation group, and the test group for evaluation. To evaluate the performance of hippocampal segmentation on the IXI dataset, we randomly selected 150 subjects from the IXI dataset for testing. To obtain the corresponding manual annotation and to reduce the workload, we first obtained the annotation with the FSL-FIRST (Patenaude et al., 2011) as the initial segmentation and then manually corrected the segmentation results to serve as the correct manual annotation.

2.3.2. Evaluation metrics

To evaluate the hippocampal segmentation results against expert manual annotation, quantitative measurements of Dice similarity coefficient (DSC), Jaccard, Precision, Recall (shown in Equation 10), Hausdorff distance, and 95*th* percentile of the distance (95% HD; shown in Equations 11, 12) were used: all of

which are standard metrics and are defined as:

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP},$$
$$DSC = \frac{2TP}{2TP + FP + FN}, \qquad Jaccard = \frac{TP}{TP + FP + FN} \tag{10}$$

$$h_{95}(X, Y) = K^{95}_{x \in X} \min_{y \in Y} |x - y|,$$
$$HD_{95}(X, Y) = \max\left(h_{95}(X, Y), h_{95}(Y, X)\right) \tag{11}$$

$$HD(X, Y) = \max(\max_{x \in X} \min_{y \in Y} |x - y|, \max_{y \in Y} \min_{x \in X} |y - x|) \tag{12}$$

In it, *TP* denotes the true positive, which means the predicted voxel coincides with the ground-truth; *FP* denotes the false positive, which means the predicted voxel falls outsides the annotation region of ground-truth; *FN* denotes the false negatives, which means the predicted background voxel is inside the ground-truth. $h_{95}(X, Y)$ is the 95*th* ranked minimum Euclidean distance between boundary points in X and Y. While DSC captures a volumetric-overlapping between the segmentations and the reference standard, the HD and 95*th* HD measure the point-wise distance.

2.3.3. Network training

All experiments were implemented in Python3.7 Pytorch backend (version 1.9) and trained on an NVIDIA RTX A6000 graphics card with 48GB of memory. In the training phase, all the network architecture was trained on the EADC-ADNI dataset with hippocampal annotation for 2,000 epochs with a batch size of 8. In the discriminator network, the convolutional kernel size is set as 4 and the stride is set as 2. To balance the training loss λ_{ce} and λ_{dice}, the parameter λ_1 is set as 0.1. And following (Tsai et al., 2018), λ_{adv} is set as 0.001. For optimization, the Adam optimizer was adopted with a learning rate of 10^{-4} for gradient update. As the hippocampus occupied only a small region in the brain scan images, to focus on the local feature around the hippocampus, each labeled MRI volume was randomly cropped to 64×64×64 voxels for model input as mentioned in Tian et al. (2021).

The sampled subjects from these two datasets are acquired from the different MR scanners at different study sites. This results in different voxel spacings, directions, and intensity ranges. Hence, before training, all the images were resampled to an isotropic voxel spacing of 1 mm/pixel according to the subjects from the EADC-ADNI dataset using the SimpleITK toolkit.

2.3.4. Examining group differences across age groups

To quantitatively measure the dentation within difference groups, we investigated the above two measurements (frequency and amplitude) in each age group and made comparisons across the groups. Finally, for quantitative analysis, we tested for statistically significant differences among age groups by performing student *t*-tests.

3. Results

In this section, we present and compare the hippocampal segmentation results of various methods on two datasets. Furthermore, to verify the robustness of the method, we also apply the framework to 7T MR scans to obtain fine segmentation results. After that, based on the proposed fine-scale segmentation method, we performed the fine-scale hippocampal morphometry study on a group of 552 healthy subjects.

Section 3.1 shows the training and validation results of different segmentation methods on the EADC-ADNI dataset. The results of the proposed fine-scale segmentation algorithm are presented in Section 3.2. Section 3.3 shows the obtained segmentation results based on the proposed framework and manual annotation in 7T scans. Finally, we used the fine-scale hippocampal segmentation from Section 3.2 to perform morphological analysis of the hippocampus of healthy subjects across different age groups, in Section 3.4.

3.1. Segmentation evaluation of EADC-ADNI dataset

3.1.1. Segmentation results at native resolution

In this section, we demonstrate the comparison between the proposed method and some of the SOTA segmentation algorithms on the EADC-ADNI dataset at the native image resolution. The results of ablation experiments of the proposed framework are also presented.

Figure 6 is a visualization of the algorithm results and manual annotations. It shows that the HippoDeep algorithm is missing some parts of the hippocampal head. This may be caused by the fact that the amygdala is sharing a very similar appearance with the hippocampus. As a result, a conservative algorithm would try to avoid leaking into the amygdala region, resulting in slight under segmentation of the hippocampal head. Along a similar vein, the HippMapp3r algorithm is missing some parts of the subiculum. The other algorithms are giving relatively satisfying results, except that the CoTr is missing bits of the CA3.

Table 2 shows the quantitative analysis results of the hippocampal predictions of various methods compared with the original labels of the EADC-ADNI dataset.

First, comparing the comparison of the segmentation results obtained from different segmentation algorithms, the proposed framework has a maximum improvement of 6.8% in DSC metrics (left hippocampus, compared to HippMapp3r) and a minimum improvement of 6.3% (left hippocampus, compared to HippMapp3r). HippMapp3r has the highest average precision. But this may be because it is more likely to under-segment the hippocampus, which is consistent with the visual appearance in the bottom panel in Figure 6. It is noticed that the HD95 and HD of the CoTr framework are better than those of the proposed method. However, CoTr sometimes suffers under-segmentation in the hippocampus region with lower DSC performance than that of ours in the experiment.

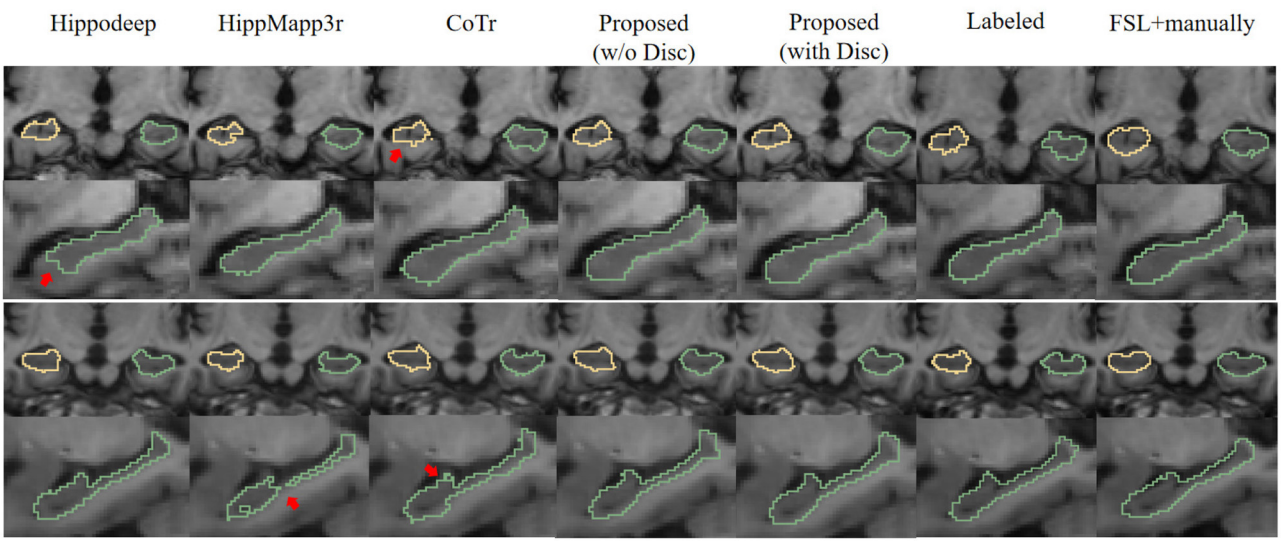


FIGURE 6
Visual contour and comparison of different segmentation methods on two examples subjects. The first and second rows are from the first subject, and the third and fourth rows are from the second subject. The first and third rows display the coronal slices of the left and right hippocampus, while the second and fourth rows display segmentation in sagittal slices. Manually traced ground truth is displayed in the penultimate column on the right. Annotation getting from FSL and manual correction results are shown in the right most column. The reason for including this "FSL+manually" column is to enable the consistent comparison of the dataset without too many manual annotations: see text in Section 3.1.2 for details. The red arrows indicate the over- or under-segmentation error.

TABLE 2 Performance comparison of segmentation results of EADC-ADNI dataset in native resolution using different methods.

Framework	Left_Right	DSC↑	Jaccard↑	Recall↑	Precision↑	HD95 (mm)↓	HD (mm)↓
Hippodeep (Thyreau et al., 2018)	Right	0.827 ± 0.027	0.706 ± 0.038	0.792 ± 0.052	0.868 ± 0.016	1.63 ± 0.49	4.23 ± 1.08
	Left	0.831 ± 0.024	0.712 ± 0.034	0.799 ± 0.052	0.87 ± 0.026	1.70 ± 0.45	4.08 ± 1.15
HippMapp3r (Goubran et al., 2020)	Right	0.808 ± 0.012	0.678 ± 0.017	0.702 ± 0.02	**0.953 ± 0.012**	1.89 ± 0.26	4.34 ± 0.88
	Left	0.812 ± 0.018	0.684 ± 0.026	0.71 ± 0.029	**0.949 ± 0.012**	1.63 ± 0.27	3.59 ± 0.61
CoTr (Xie et al., 2021)	Right	0.862 ± 0.012	0.757 ± 0.018	0.878 ± 0.027	0.847 ± 0.017	1.47 ± 0.17	3.36 ± 0.51
	Left	0.862 ± 0.01	0.758 ± 0.015	0.892 ± 0.028	0.835 ± 0.024	1.44 ± 0.09	3.42 ± 0.39
Proposed (w/o Disc)	Right	0.873 ± 0.008	0.774 ± 0.013	0.893 ± 0.015	0.854 ± 0.019	1.41 ± 0.01	**3.32 ± 0.59**
	Left	0.873 ± 0.011	0.774 ± 0.017	0.907 ± 0.02	0.842 ± 0.022	1.41 ± 0.16	**3.09 ± 0.44**
Proposed (w Disc)	Right	**0.876 ± 0.008**	**0.779 ± 0.012**	**0.906 ± 0.018**	0.848 ± 0.018	**1.41 ± 0.01**	3.38 ± 0.63
	Left	**0.875 ± 0.012**	**0.778 ± 0.019**	**0.910 ± 0.024**	0.844 ± 0.027	**1.37 ± 0.20**	3.35 ± 0.69

The upward arrow indicates that higher values are better and the downward arrow indicates that lower is better. The best results are marked in bold.

Second, to verify the effectiveness of each component of the proposed framework, two ablation experiments are conducted and the results are shown in Table 2. (1) The proposed framework without the discriminator is improved based on CoTr. Compared to CoTr, our proposed framework without the discriminator part achieved a higher overlap evaluation score and lower segmentation error around the edge of the hippocampus (increased by 0.3 mm in HD metric on the left). (2) Comparison of the models with and without the discriminator: we found that the former achieves the best performance in most evaluation metrics. In summary, these two ablation experiments demonstrate the effectiveness of our proposed framework for improving the segmentation performance of the hippocampus at the native resolution.

3.1.2. Constructing consistent hippocampus annotations between two datasets

Although the EADC-ADNI dataset contains the 3D manual annotation, the IXI dataset, unfortunately, does not. Therefore, to conduct a consistent comparison between the two datasets, we have to create a consistent reference for both. Since directly labeling the hippocampus would be time-consuming, following (Liu et al., 2020), we used the results from FSL-FIRST (Patenaude et al., 2011) as the initial segmentation and make manual corrections afterward. This created consistent 3D reference annotations for the two datasets. Next, they are collectively named FSL+manually labeled.

Table 3 presents the quantitative analysis results of the hippocampal segmentation of various methods compared with the

FSL+manually label of the EADC-ADNI dataset. As can be seen, the proposed method can obtain the highest DSC metric. Moreover, the highest segmentation accuracy was obtained for both sides' hippocampi measured by the Jaccard, Recall, and HD95 metrics, similar to the case in Table 2. Although the DSC of our proposed method in Table 3 are approximately 3% lower than those in Table 2, this reduction is also observed in the other segmentation models. The performance of these segmentation models on the other metrics also shows a consistent change from Table 2 to Table 3, with Recall decreasing and Precision increasing. Moreover, the evaluation results in Table 2 are higher than those in Table 3 because in Table 2 those annotations for validation and training are drawn manually from the same group of annotation experts. Therefore, a reduction in the evaluation metrics is caused by the variability of the two different manual annotations.

It is noteworthy that all the above segmentation results were obtained at the native image resolution. Although they all achieve quite high evaluation metrics, as can be seen in Figure 7C, at the sub-pixel level, the bumpy structure can be clearly seen. However, on the reconstructed surface, the staircase appearance does not indicate the correct local morphology. This is because all the methods above are based on training annotations. However, under the native resolution, even manual annotation can not correctly characterize the bumps. Because the function space limited by the native resolution determines the morphological characterization capability. Such a surface space limitation should be addressed, for the segmentation to correctly characterize the bumps/dentations. To solve this problem, we need the help of fine-scale segmentation to get the fine-scale annotations.

3.1.3. Fine-scale segmentation for EADC-ADNI dataset

The fine-scale segmentation results for the EADC-ADNI dataset are shown in Figure 7E. It can be seen that the smooth curves accurately delineate the bumpy hippocampal structure. The last row in Table 3 shows the quantitative analysis results of the fine-scale segmentation results compared with the FSL+manually label of the EADC-ADNI dataset. As seen in that row, the value of the DSC metric increased by about 2% and the precision value also increased compared to Table 3. This may be due to the curvature regularization at a much higher resolution. While it successfully regulates the surface evolution from generating singularities, inevitably, it will shrink the total volume slightly and result in a more conservative segmentation.

It can also be seen from the quantitative results in the last row of Table 3, the fine-scale segmentation results have improvements in the DSC and Jaccard similarity coefficients compared to the native-resolution-based results. This is consistent with the visualization results shown in Figure 7, since these two metrics are volume-based evaluation metrics. Nevertheless, it still cannot reflect the changes in dentation segmentation significantly. Because fine-scale segmentation is reflected more on improving the accuracy on the boundary, rather than on the volumetric measurement. Therefore, surface-distance-based metrics, such as HD, can better demonstrate improvements in fine-scale segmentation. However, a full 3D delineation across the thousands of slices at a much

TABLE 3 Quantitative comparison of segmentation results obtained by different segmentation methods with results from FSL+manual annotation on EADC-ADNI dataset.

Framework	Left_Right	DSC↑	Jaccard↑	Recall↑	Precision↑	HD95 (mm)↓	HD (mm)↓	2D HD (mm)
Hippodeep (Thyreau et al., 2018)	Right	0.800 ± 0.033	0.668 ± 0.044	0.715 ± 0.062	0.918 ± 0.054	2.13 ± 0.39	3.92 ± 0.76	2.12 ± 0.64
	Left	0.812 ± 0.025	0.684 ± 0.035	0.733 ± 0.05	0.915 ± 0.045	1.91 ± 0.31	3.82 ± 0.70	1.91 ± 0.43
HippMapp3r (Goubran et al., 2020)	Right	0.750 ± 0.026	0.60 ± 0.033	0.612 ± 0.04	**0.970 ± 0.022**	2.38 ± 0.28	4.49 ± 0.74	2.41 ± 0.52
	Left	0.755 ± 0.022	0.607 ± 0.028	0.624 ± 0.031	**0.956 ± 0.026**	2.27 ± 0.21	4.31 ± 1.01	2.14 ± 0.49
CoTr (Xie et al., 2021)	Right	0.820 ± 0.017	0.695 ± 0.024	0.752 ± 0.036	0.904 ± 0.029	1.99 ± 0.19	3.86 ± 0.42	1.65 ± 0.44
	Left	0.820 ± 0.019	0.696 ± 0.027	0.779 ± 0.025	0.868 ± 0.04	1.95 ± 0.26	**3.84 ± 0.83**	1.78 ± 0.55
Proposed (w/o Disc)	Right	0.834 ± 0.012	0.715 ± 0.017	0.79 ± 0.026	0.884 ± 0.034	1.99 ± 0.18	4.12 ± 0.65	1.52 ± 0.38
	Left	0.827 ± 0.021	0.705 ± 0.03	0.803 ± 0.021	0.854 ± 0.04	1.94 ± 0.39	4.06 ± 0.95	1.70 ± 0.53
Proposed (w Disc)	Right	**0.847 ± 0.013**	**0.734 ± 0.019**	**0.810 ± 0.029**	0.888 ± 0.029	**1.81 ± 0.28**	**3.83 ± 0.66**	**1.24 ± 0.31**
	Left	**0.837 ± 0.020**	**0.72 ± 0.029**	**0.813 ± 0.026**	0.864 ± 0.04	**1.85 ± 0.42**	3.89 ± 1.55	**1.55 ± 0.56**
Proposed (SR)	Right	0.866 ± 0.019	0.765 ± 0.029	0.823 ± 0.044	0.918 ± 0.027	1.49 ± 0.22	3.41 ± 0.84	1.16 ± 0.32
	Left	0.863 ± 0.016	0.759 ± 0.024	0.808 ± 0.035	0.928 ± 0.033	1.49 ± 0.31	3.47 ± 1.64	0.87 ± 0.19

SR indicates that the comparison is on the fine scale. The best results of each metric at native resolution are marked in bold.

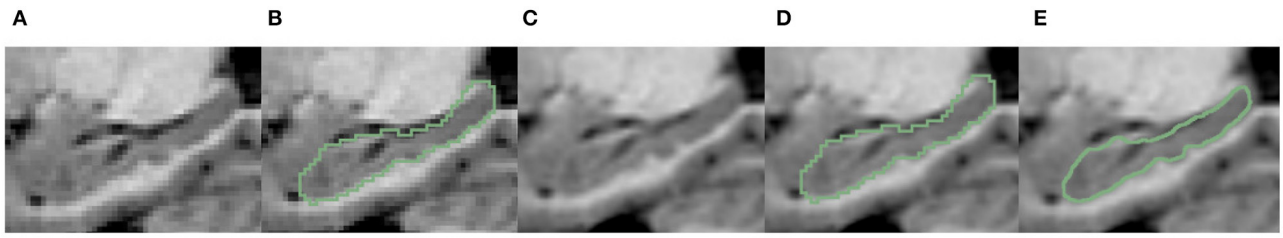


FIGURE 7
Visualization comparison of hippocampal segmentation results obtained at native image resolution and fine scale in EADC-ADNI dataset. **(A)** Shows the MR image of hippocampus at native resolution; **(B)** shows the segmentation results of hippocampus at native resolution in green contour line; **(C)** shows the MR image after interpolation at fine scale; **(D)** shows the native resolution hippocampus segmentation overlaid on the fine-scale MR image after interpolation; **(E)** shows the fine-scale hippocampus segmentation overlaid on the fine-scale MR image.

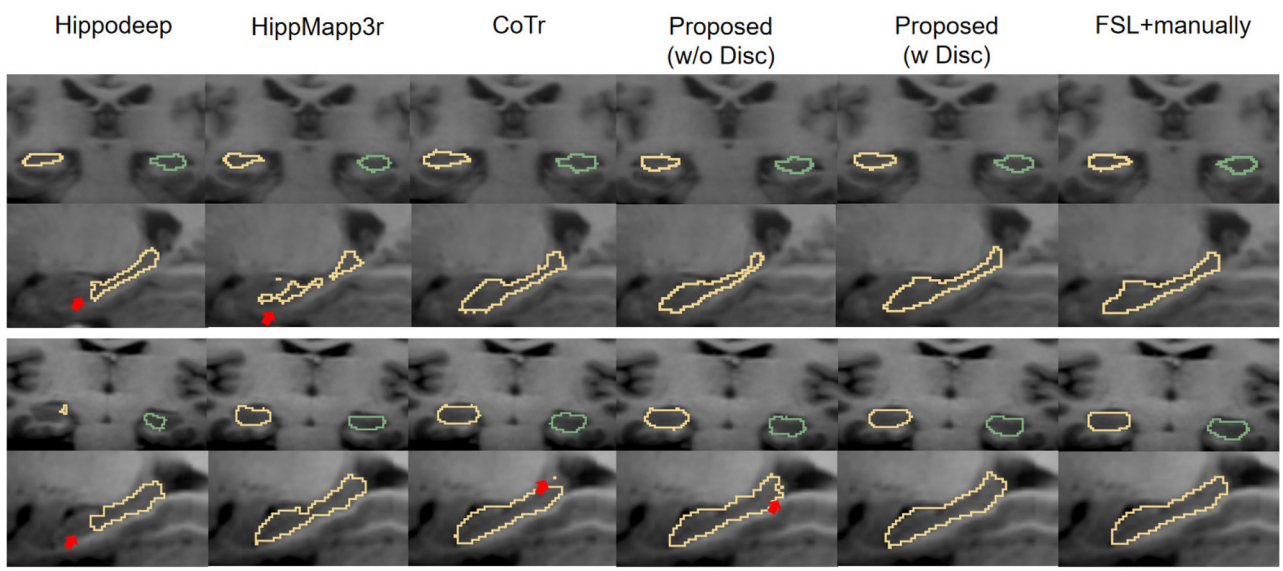


FIGURE 8
Visual contour and comparison of different segmentation methods on two example subjects from the IXI dataset. The first and second rows are from the first subject, and the third and fourth rows are from the second subject. Yellow and green outlines indicate left and right hippocampus segmentation. Manually traced results are displayed in the last column on the right. The red arrows indicate the over- or under-segmentation error.

higher resolution in a consistent manner is extremely tedious, if not impossible. To solve that dilemma, we use two-dimensional Hausdorff distance (2D HD) at certain characteristic slices for quantitative evaluation.

The characteristic sagittal slice that best reflects the bumpy features on the surface of the hippocampus is selected at fine-scale resolution. Then, the boundaries of hippocampal dentation are outlined manually. After that, for comparison, the fine-scale hippocampal annotation on the same slice was extracted. The Hausdorff distance between its boundary and the manually drawn contour was calculated. The results are shown in Table 3 (the "2D HD" column). As can be seen, the segmentation of dentation at fine scale is significantly improved reflected by this evaluation metric. This indicates that the fine morphology is better captured at such a fine-scale resolution than that at the native image resolution. It is also consistent with the visualization results shown in Figure 7, with more accurate dentation annotation at the fine scale.

3.2. Segmentation evaluation of IXI dataset

The ultimate goal of performing fine-scale segmentation is to extract the detailed hippocampal morphology, which can be used for cross-sectional and/or longitudinal comparisons among different groups of subjects. With such a goal in mind, while the EADC-ADNI dataset has manual annotation at the native image resolution, we have to deploy the algorithm to a much larger set for the morphometry. Unfortunately, such a dataset of 552 healthy subjects does not have a complete reference annotation.

In this section, the fine-scale segmentation is carried out and evaluated on such a much larger dataset.

3.2.1. Segmentation comparison at native resolution

As mentioned above, since the IXI dataset does not provide annotations of the hippocampus, following (Liu et al., 2020),

TABLE 4 Quantitative comparison of segmentation results obtained by different segmentation methods with results from FSL+manual annotation on IXI dataset.

Framework	Left_Right	DSC↑	Jaccard↑	Recall↑	Precision↑	HD95 (mm)↓	HD (mm)↓	2D HD (mm)↓
Hippodeep (Thyreau et al., 2018)	Right	0.835 ± 0.070	0.721 ± 0.077	0.814 ± 0.076	0.861 ± 0.069	1.92 ± 1.64	4.47 ± 1.86	1.72 ± 1.62
	Left	0.835 ± 0.067	0.721 ± 0.073	0.806 ± 0.076	0.873 ± 0.065	1.89 ± 1.69	4.32 ± 1.97	1.75 ± 1.96
HippMapp3r (Goubran et al., 2020)	Right	0.812 ± 0.053	0.686 ± 0.062	0.735 ± 0.059	**0.911 ± 0.071**	1.91 ± 0.65	4.34 ± 1.21	1.91 ± 0.54
	Left	0.807 ± 0.050	0.678 ± 0.056	0.720 ± 0.061	**0.922 ± 0.065**	1.95 ± 0.51	4.16 ± 1.11	1.74 ± 0.59
CoTr (Xie et al., 2021)	Right	0.846 ± 0.017	0.734 ± 0.025	0.821 ± 0.028	0.876 ± 0.039	1.68 ± 0.32	4.08 ± 0.52	**1.67 ± 0.95**
	Left	0.850 ± 0.019	0.740 ± 0.028	0.824 ± 0.027	0.880 ± 0.038	1.69 ± 0.30	4.02 ± 0.61	1.56 ± 0.87
Proposed (w/o Disc)	Right	0.850 ± 0.028	0.741 ± 0.04	0.806 ± 0.045	0.902 ± 0.039	1.69 ± 0.34	**3.86 ± 0.62**	2.03 ± 1.41
	Left	0.854 ± 0.030	0.746 ± 0.043	0.806 ± 0.042	0.909 ± 0.040	1.73 ± 0.36	3.78 ± 0.66	1.93 ± 1.42
Proposed (w Disc)	Right	**0.865 ± 0.020**	**0.763 ± 0.030**	**0.841 ± 0.036**	0.894 ± 0.039	**1.60 ± 0.31**	3.98 ± 0.64	1.78 ± 1.22
	Left	**0.867 ± 0.021**	**0.765 ± 0.033**	**0.843 ± 0.031**	0.894 ± 0.041	**1.63 ± 0.33**	**3.70 ± 0.70**	**1.52 ± 1.35**
Proposed (SR)	Right	0.866 ± 0.022	0.765 ± 0.033	0.835 ± 0.039	0.903 ± 0.045	1.40 ± 0.29	3.17 ± 0.50	1.24 ± 0.96
	Left	0.865 ± 0.023	0.763 ± 0.035	0.837 ± 0.036	0.897 ± 0.04	1.50 ± 0.28	3.08 ± 0.55	1.25 ± 1.39

SR indicates that the comparison is on a fine scale. The best results of each metric at native resolution are marked in bold.

we obtained the hippocampus segmentation with the help of FSL software and manual correction as the reference ground truth in IXI dataset. Then, it is used to evaluate the segmentation performance of the proposed framework and other SOTA algorithms on the IXI dataset. The experiments described next are based on the randomly selected 150 sample subjects from the IXI dataset as the research objects.

Figure 8 shows the visual comparison of the segmentation results given by different segmentation models at the native image resolution. It shows that Hippodeep fails to capture the head of the hippocampus. Likewise, HippMapp3r does not perform well in the same example subject. The output of CoTr is incomplete but unlike the previous examples, it omits the caudal part of the hippocampus. Improved from CoTr, our proposed basic segmentation framework (without discriminator) is not constrained by the input size and can output more complete segmentation results.

The reason for the under-segmentation of the IXI dataset by these above methods may be that they were not directly trained by the annotations of the IXI dataset, and the distribution of hippocampal samples on the IXI dataset has not been seen before. In contrast, the proposed framework with discriminator can utilize the new images to adaptively improve the generalization ability of the model. Therefore, the proposed framework performs well in these samples, and its output is visually closer to the ground-truth annotations.

Table 4 shows the quantitative results using different methods on the 150 sample objects of IXI dataset. Combining the visualization results in Figure 8 and the quantitative analysis results in Table 4, we have the following findings.

First, the Dice score of Hippodeep is higher than that of HippMapp3r. However, Hippodeep has certain unsatisfactory segmentation as shown in Figure 8, resulting in a larger standard deviation. Additionally, the precision of the segmentation of HippMapp3r is the highest among all methods. As can be seen in Figure 8, the output of HippMapp3r was more likely to be located within the hippocampal region. Therefore, the segmentation results of HippMapp3r have lower false positives and thus the highest precision. As for the evaluation results based on HD and 2D HD distance metrics, the proposed model achieves the best results on the left hippocampus than that on the right. Furthermore, compared with other frameworks, our proposed segmentation model achieves the best results on the other evaluation metrics. In particular, it is approximately 5% higher than the evaluation metric of HippMapp3r on DSC. To sum up, according to the quantitative and qualitative comparison results, the proposed framework's output is more satisfactory than hippocampal segmentation results on the IXI dataset.

Moreover, another advantage over state-of-the-art methods (Hippodeep and HippMapp3) is that only 30 subjects from another dataset (EADC-ADNI dataset) are used for training, while the testing set in this study involves a different cohort with 150 sample subjects. This shows the generalization capability of the proposed method.

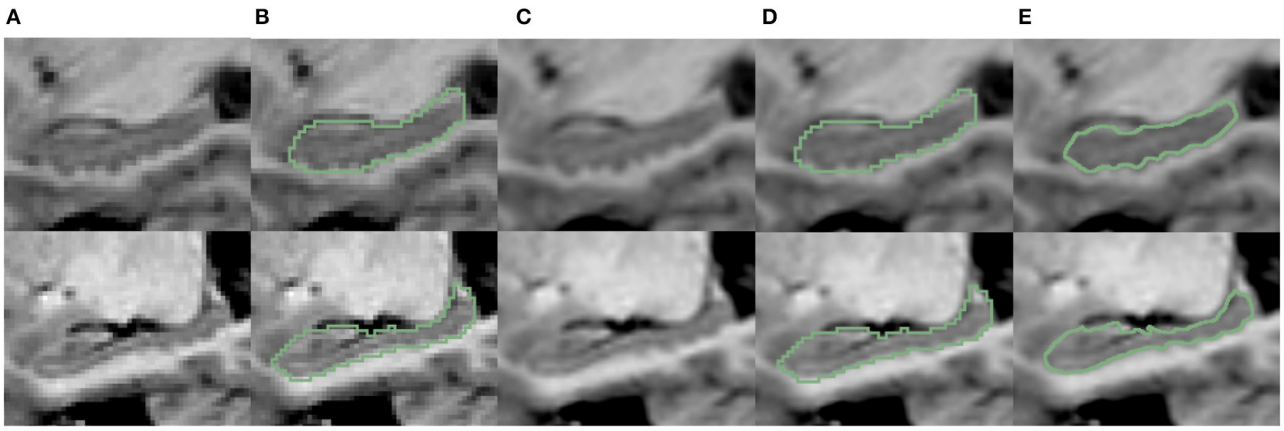


FIGURE 9
Visualization comparison of hippocampal segmentations at native image resolution vs. fine scale in IXI dataset. **(A)** The MR image of the hippocampus at native resolution. **(B)** The segmentation of hippocampus at native resolution; **(C)** shows the MR images after interpolation; **(D)** shows the native resolution hippocampus segmentation overlaid on the fine-scale MR image after interpolation; **(E)** shows the fine-scale hippocampus segmentation overlaid on the fine-scale MR image.

3.2.2. Evaluation of fine-scale segmentation

The above results are based on the native image resolution. To accurately analyze the change of dentation, we applied the fine-scale segmentation method to obtain the segmentation and compare them with manual annotation at native resolution and fine-scale, visually shown in Figure 9. Similar to Figures 7B, D, it can be seen from Figures 9B, D that the segmentation model can only output rough stepped edges at the native image resolution, but fails to capture the edges of the hippocampal dentation. Further, compared with the segmentation results in Figures 9B, D, the Figure 9E shows that the proposed method can better capture the dentation structure of the hippocampus, resulting in finer segmentation results.

The last row of Table 4 shows the quantitative analysis of the fine-scale segmentation results. Compared to the performance at native resolution, overlap-based metrics (i.e., DSC, Jaccrd, Recall, and Precision) did not show significant changes. But the distance-based metrics (i.e., HD95, and HD) decrease significantly. Among them, the HD metric can be lowered by up to 0.8 mm (about 20% for the right hippocampus). To measure the improvement of boundary fineness by fine-scale segmentation methods, we also focused on the results for 2D HD. There is also a 0.5 mm (about 28%) drop for the right hippocampus in the 2D HD metric for dentation.

Combining the overlap-based and distance-based metrics shows that the fine-scale segmentation algorithm does not have a great impact on the segmentation accuracy of the overlapped region of the hippocampus. Instead, the algorithm can change the edges of the segmented objects, thereby improving the accuracy of the dentate segmentation.

Finally, we applied the proposed method and obtained fine segmentation results for all 552 case samples based on the IXI dataset. Some of the visual results are shown in Figure 10. The bumps on the inferior side of the hippocampus can be captured in the presented samples and they look different among different age groups. However, the "bumpiness" across different age groups can not be easily assessed by eye, and we need to use quantitative metrics to do so. This is subject to the topic of the next section.

Since the most important bumpy dentation information can be observed in the 2D slices, we segmented the volumetric data and validated the accuracy in 2D.

3.3. Validation with 7T MR images

To validate our segmentation accuracy on 1 mm/pixel MR scans against high-resolution 7T MR scans, we mimic lower resolution images using 7T MR scans, applied the proposed method, and validated the result against the manual contour of the inferior surface slices which shows the most prominent bumps at high resolution. For analysis, we selected three samples from the dataset provided by Alkemade et al. (2020) for testing. Since the most important bumpy dentation information can be observed in the 2D slices, we segmented the volumetric data and validated the accuracy in 2D. First, we down-sampled the 7T MR images with a resolution of 0.641 to 1 mm. We then applied the proposed segmentation framework to obtain fine-scale segmentations. The segmentation results were presented in Figure 11. Observing Figures 11B, C, it can be seen that the native segmentation results at 1 mm resolution in Figure 11B cannot accurately capture the boundaries of hippocampus dentation well, though it can be observed in Figure 11C, the original high-resolution slices. Conversely, the fine-scale segmentation results shown in Figure 11E, obtained with the proposed framework, exhibited a high degree of consistency with the manual annotation in Figure 11D. Quantitatively, the DSC obtained on the three slices were 0.892, 0.897, and 0.861, respectively, while the 95th percentile of the Hausdorff distance measured 0.641, 0.640, and 0.906 mm. These results demonstrate that our fine-scale segmentation approach yields accurate outcomes, which closely align with the true 7T segmentation results obtained from high resolution MR scans.

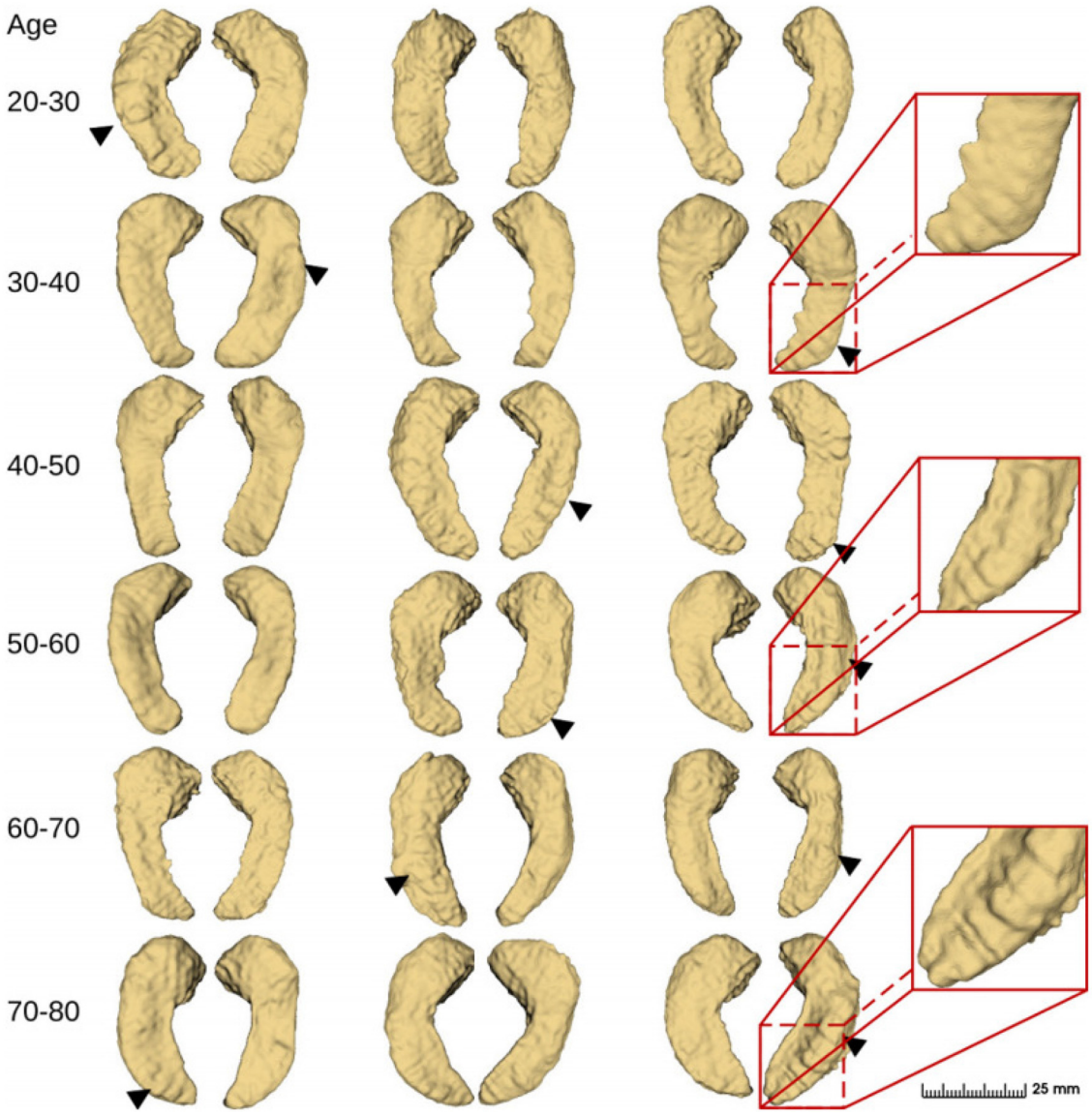


FIGURE 10
Fine-scale segmentation by the proposed method in different age groups from the IXI dataset (2018). Black triangles indicate hippocampal dentation.

3.4. Shape analysis of the hippocampal dentation in fine scale

Sections 3.2 and 3.1 above validated the segmentation accuracy of the proposed framework. The ultimate goal of the present work is to quantitatively analyze the fine-scale dentation feature underneath the hippocampus using the methods in Section 2.2. First, we quantitatively validated the hippocampal dentation analysis method in Section 2.2 on some simulated shapes in Section 3.4.1. We then applied the validated methods on the real fine-scale segmentation results in Section 3.4.2 and identified the trends of hippocampal dentation through different age groups.

3.4.1. Quantitative validation of hippocampal dentation analysis on simulated shapes

In this section, we quantitatively evaluate the hippocampal dentation analysis method used in Section 2.2 and show its accuracy in capturing the magnitude and frequency of dentation patterns.

Since there is no established ground truth for the measurements of the dentation patterns, it would be difficult to evaluate the accuracy if we directly apply the methods to the real anatomical structures. As a result, following the ideas in Gao et al. (2014), we generated a series of simulated shapes, with known varying dentation patterns in their magnitudes and frequencies. Such shapes are then used to evaluate the analysis method.

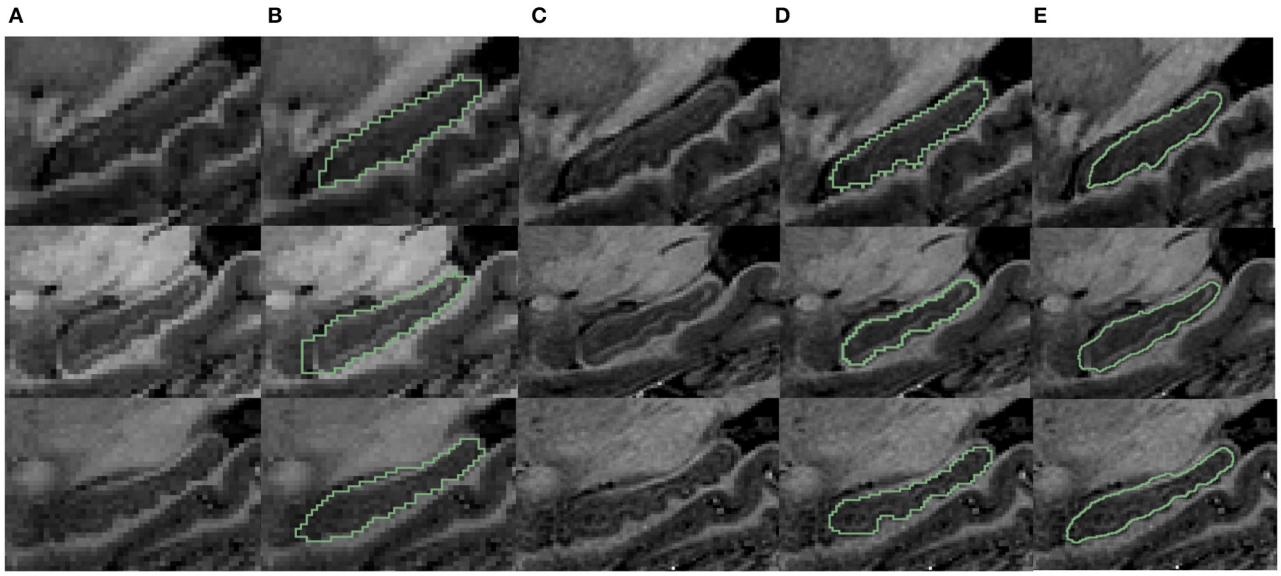


FIGURE 11
Visualization comparison of hippocampal segmentations in 7T MR scans. **(A)** The MR image of the hippocampus at resampled 1 mm resolution. **(B)** The obtained segmentation of the hippocampus at 1 mm resolution overlaid on **(A)**; **(C)** shows the MR images at native 0.641 mm resolution from 7T scans; **(D)** shows the manual annotation overlaid on the 7T MR images; **(E)** shows the fine-scale hippocampus segmentation with resolution 0.2 mm overlaid on the native 7T MR image with resolution of 0.641 mm.

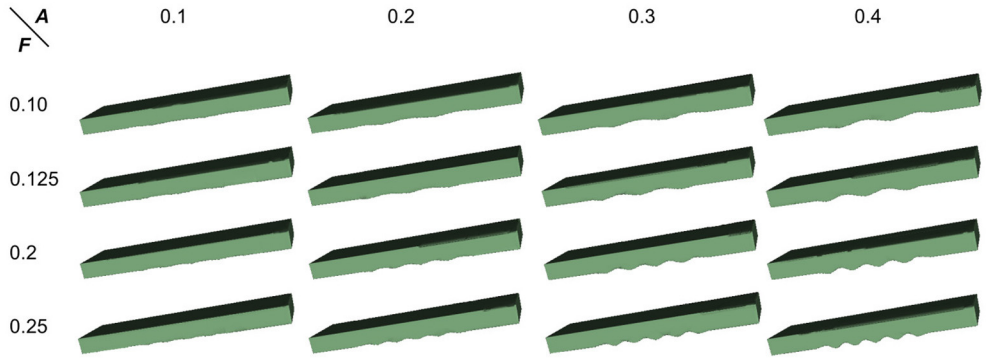


FIGURE 12
Simulated 3D hippocampal dentation. There are 16 combinations of dentation simulations with different amplitudes (***A***, ranging from 0.1 to 0.4 mm) and frequencies (***F***, ranging from 0.1 to 0.25 bump/mm).

To proceed, we simulated the dentations with dentation amplitudes and frequency ranges concerning (ten Hove and Poppenk, 2020) on a cuboid, as shown in Figure 12. Then, following the proposed shape analysis method, we extracted both the amplitudes and the frequencies. Corresponding to the simulation example in Figure 12, the visual examples of the results obtained from dimensionality reduction and curve fitting are shown in Figure 13. Based on the fitted curve, we computed the dentation frequencies and amplitude. Finally, the computed amplitudes and frequencies are compared with their ground truth. The evaluation results are shown in Table 5.

As can be seen from Table 5, the fitting error of the frequencies is below 1%. This is partially because the shapes are quite ideal. However, the error of the amplitude detection is larger, with a maximum error of about 0.019 mm. It can also be observed from Table 5 that the amplitudes are often time underestimated. However, when the amplitude is larger than 0.2 mm, the statistical error in Table 5 is greatly reduced to around 4.5%. As the amplitude increases to 0.4 mm, the fitting error is lower than 1%. Consistent with this, it can also be seen from Figure 13 that the fitted curve is almost the same as the actual curve.

The above quantitative analysis of errors shows that, based on dimensionality reduction and curve fitting, the proposed shape analysis method has relatively larger errors when the bumps are shallow. With the gradual increase of the bump amplitude, the error decreases to about 1%. After this validation of the simulated data was completed, we then applied the method to the real segmentation results.

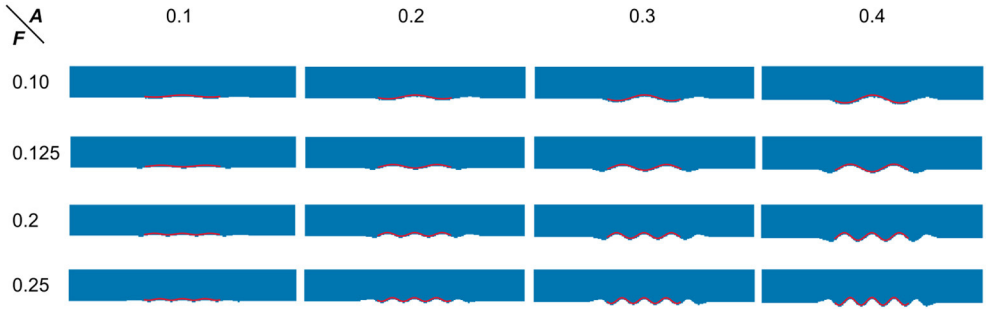


FIGURE 13
Visual examples of dimensionality reduction and curve fitting. Blue points represent the result of the 3D simulation model after dimensionality reduction. The red wavy curve represents the fitted curve.

3.4.2. Quantitative analysis of hippocampal dentation —The hippocampus is the bumpiest in people in their 40s

Utilizing the proposed segmentation method, we captured the fine-scale dentations on the IXI dataset. Shape analysis was subsequently performed on the annotated data from the IXI dataset, consisting of 552 healthy subjects, using the method described in Section 2.2. Figure 14 shows the variation in hippocampal dentation amplitude and frequency sub-stratified by age. The higher the amplitude, the higher the hippocampus dentation. Higher frequencies indicate narrower dentation in the hippocampus.

As depicted in Figure 14, the dentations under the hippocampus are most pronounced in the age group of people between 40 and 50 years old. First, there were more variations of amplitude in the 40 to 50 age group, which ranged approximately from 0.09 to 0.2 on both sides in Figures 14A, C. On the other hand, the change in frequency trended in the opposite direction to the change in amplitude but still reached its lowest point at the age of 40 to 50, and ranged from 0.11 to 0.16.

Figure 14 shows inter-group statistical analysis and the differences by two sample independent *t*-tests. The temporally aligned blocks for six groups reveal distinct ($P < 0.05$) patterns in hippocampus dentation. The most notable differences between groups were the amplitude of left hippocampus dentation (group 40–50/others).

4. Discussion and conclusion

This work has presented a complete pipeline of fine-scale hippocampus segmentation and dentation analysis. Results indicated that this is an efficient method for accurate sub-millimeter hippocampus segmentation and dentation shape variation analysis in 3T MR images in different age groups.

The proposed method addressed the two main difficulties of obtaining fine-scale annotation of the hippocampus efficiently from clinically available image data instead of ultra-high field MR scans and exploring the relationship between hippocampal longitudinal dentation and age in normal and healthy groups.

For hippocampus segmentation, the proposed algorithm based on 3D deep neural networks improved the segmentation performance and efficiency, which fulfilled the need to obtain annotation of the hippocampus of a large cohort with 552 sample subjects. Only a small sample size of 30 volumes was used for model training and hippocampus segmentation tasks. To solve the problem of the difference in the distribution of training and testing samples, we improved from the CoTr model and utilized the domain adaptation method to improve the performance of validation and testing on the second dataset. For example, the segmentation performance of the tail in the hippocampus was improved. This deep learning based semantic segmentation method provided accurate initial segmentation for the subsequent fine-scale segmentation. Furthermore, to compare the change of hippocampal segmentation results at the native resolution and the fine scale, we applied distance-based evaluation metrics. The reduction of HD and 2D HD showed that, with the help of fine-scale segmentation algorithms for morphological analysis, segmentation results could better capture the outline of the whole hippocampus and its dentation.

To the best of our knowledge, this is the first quantitative investigation of fine-scale hippocampus morphometry across a wide range of age groups. This initial study reveals noticeable patterns of shape changes. Dentation of the hippocampus is present during the initial stages of life and continues to change as the individual grows. These changes are commensurate in relative extent with the temporal structural evolvement of the hippocampus within the first few decades up to the age of 50. By contrast, the dentational region undergoes a lower rate of change, leading to a relative degree of loss in the inferior regions of the hippocampus. Although the total change rate of dentational regions presents concavity or convexity for the corresponding quantitative parameters, the reverse is not true for people in older age groups: in these individuals, with severe tissue loss, dentation has a more irregular outline.

These findings are consistent with continuous variability across the full spectrum of neurogenesis, as is increasingly being verified from the molecular structure level (Alvarez-Buylla and Lim, 2004; Lim and Alvarez-Buylla, 2016). Another study, in Wu et al. (2013), has demonstrated that the structural brain network also peaks between 40 and 50 years of age.

TABLE 5 Quantitative error analysis results between the fitting and the actual setting frequency (F) and amplitude (A).

Actual A (mm) / Actual F (bump/mm)	0.1	0.2	0.3	0.4	MAE (F)
0.1	0.082 (0.101)	0.193 (0.099)	0.297 (0.1)	0.401 (0.1)	0.001
0.125	0.082 (0.124)	0.19 (0.124)	0.299 (0.124)	0.398 (0.125)	0.001
0.2	0.082 (0.2)	0.191 (0.199)	0.295 (0.2)	0.394 (0.2)	0
0.25	0.08 (0.251)	0.189 (0.249)	0.294 (0.25)	0.396 (0.25)	0.001
MAE (A)	0.019	0.009	0.004	0.003	\

In parentheses are the fitted frequencies.

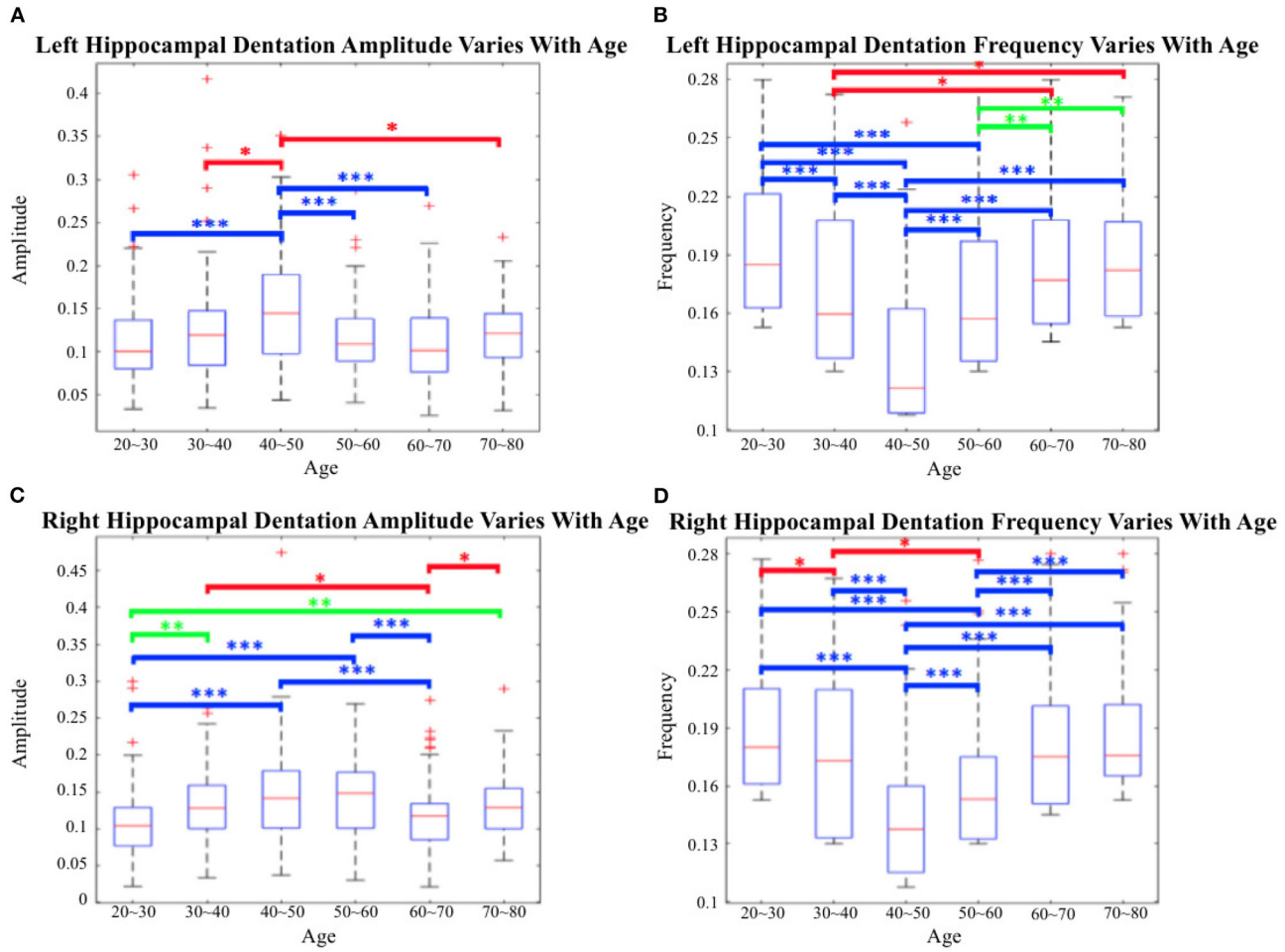


FIGURE 14
Cross-age group assessments. **(A, C)** Box-plots of left hippocampal dentation amplitudes and frequency variation with age; **(B, D)** Box-plots of right hippocampal dentation amplitudes and frequency variation with age. The bridges of different colors span between the two groups, indicating that their difference is statistically significant. *$P < 0.05$ (indicated in red); **$P < 0.005$ (indicated in green); ***$P < 0.001$ (indicated in blue).

Our findings support the idea that the temporal profiles of dentation in healthy subjects may be the consequence of neurogenesis at the specific site of brain regions. It has also been demonstrated that adults preserve neural stem cells, which produce new neurons within some restrained areas. These cell populations could be viewed as displaced and modified neuroepithelium, pockets of cells, and local signals that retain enough embryonic nature to maintain neurogenesis for life. These findings suggest a selective cortical variation that is consistent with the extent and dynamics of neurogenesis, with the most active growth happening during embryonic development, followed by continuous generation, decreasing slowly with age (Knoth et al., 2010; Sanai et al., 2011; Göritz and Frisén, 2012). However, regions of neurogenesis exhibit pathological distinctions between healthy subjects and some neurodegenerative disease patients, and these distinctions are evident throughout the course of a disease. For

example, in Huntington's disease, it has been found that postnatally generated neurons are absent in the advanced stages of disease (Zuccato et al., 2010; Walker et al., 2011). Accordingly, serious consideration should be given to which factors might result in distinctions of neurogenesis activity, and whether there is any associated phenotype, just as Huntington's disease subjects seemed to reveal a more pronounced rate of atrophy within specific regions of interest. Similarly, longitudinal model-based estimation of variations and distinct phenotypic variability of dentation compared to healthy subjects could not be neglected during some neurodegenerative conditions.

A key advantage of this work is that it develops methods for quantifying the continuous phenotypic variability of dentation, which ranges from completely absent to pronounced among healthy adults. The proposed method extracts prominent change patterns from 3D volume data, which are critical for subsequent evaluation and to provide an effective feature expression. Compared to previous cross-sectional studies (Beattie et al., 2017), our work dispenses with a burdensome and subjective visual rating process. The non-linear fitting model provides two parameters—amplitude and frequency—to permit quantitative analysis of variation. However, the framework can only integrate dentation contour to a sinusoidal locus where the modeled average rate of change of mass data can support the model-based estimation.

The amplitudes and frequencies we measured are smaller than those found in ten Hove and Poppenk (2020). There may be several reasons for this: first, the IXI dataset we used are vanilla T1-MPRAGE sequence, which are not designed to highlight the hippocampal dentations. Second, in Section 2.2, we used a linear projection to map the 3D shape to 2D curves, which were later fitted with a sinusoidal function. In this process, the direction of projection may not be perfectly aligned with the ridge of the dentation due to its non-planner/non-linear nature. Furthermore, the inferior surface of the hippocampus is not a flat plane. The combination of these factors could decrease the amplitude of dentation in the 2D view and subsequent sinusoidal fitting. Further research investigating better bump extraction and parameterization approaches, such as the principal curve analysis and/or machine learning based approaches, is ongoing.

As the first systematic temporal study of hippocampal fine-scale dentation that includes analyses of 3T clinical data and comprehensive neuroanatomical measures, a few limitations to the present work have to be noted. Even though we validated that the dentations found in the proposed method are not interpolation artifacts, as seen in Chang et al. (2018), it is preferable to obtain paired 3T and 7T datasets for further validation of the dentation delineation. Moreover, the imaging data in this study were acquired from public databases. To enhance the robustness and generalization of the estimation model, promoting more studies spanning different databases from different sites and large-scale analysis to integrate these data is required. For instance, after the axis extraction, the non-linear fitting is susceptible to the local minimum. Even though the simulated annealing can ameliorate this situation to some extent, further improvements in dentation feature extraction need to be undertaken in future research.

In addition to the healthy subjects studied here, future directions of this research should also explore the potential diagnostic and prognostic utility of patterns of dentation in disease states, as well as serving as an outcome measure for interventions, such as epilepsy, Alzheimer's disease, and schizophrenia.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

QY and SC: methodology, software, and writing the manuscript. GC, RC, CH, and TR: writing the manuscript. XY and NZ: methodology. YG: conceptualization of this study, methodology, and writing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the Key-Area Research and Development Program of Guangdong Province (Grant Number 2021B0101420005), the Key Technology Development Program of Shenzhen (Grant Number JSGG20210713091811036), the Department of Education of Guangdong Province (Grant Number 2017KZDXM072), the National Natural Science Foundation of China (Grant Number 61601302), the Shenzhen Key Laboratory Foundation (Grant Number ZDSYS20200811143757022), Shenzhen Peacock Plan (Grant Number KQTD2016053112051497), and the SZU Top Ranking Project (Grant Number 86000000210).

Acknowledgments

This study used samples from the EADC-ADNI Harmonized Protocol project, IXI, and AHEAD datasets. We would like to express our gratitude for their contribution to this research.

Conflict of interest

CH declares no external support related to this work.

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

References

Alkemade, A., Mulder, M. J., Groot, J. M., Isaacs, B. R., van Berendonk, N., Lute, N., et al. (2020). The Amsterdam ultra-high field adult lifespan database (ahead): a freely available multimodal 7 Tesla submillimeter magnetic resonance imaging database. *Neuroimage* 221, 117200. doi: 10.1016/j.neuroimage.2020.117200

Alvarez-Buylla, A., and Lim, D. A. (2004). For the long run: maintaining germinal niches in the adult brain. *Neuron* 41, 683–686. doi: 10.1016/S0896-6273(04)00111-4

Apostolova, L. G., Dinov, I. D., Dutton, R. A., Hayashi, K. M., Toga, A. W., Cummings, J. L., et al. (2006). 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. *Brain* 129, 2867–2873. doi: 10.1093/brain/awl274

Apostolova, L. G., Zarow, C., Biado, K., Hurtz, S., Boccardi, M., Somme, J., et al. (2015). Relationship between hippocampal atrophy and neuropathology markers: a 7T MRI validation study of the EADC-ADNI harmonized hippocampal segmentation protocol. *Alzheimers Dement.* 11, 139–150. doi: 10.1016/j.jalz.2015.01.001

Arslan, O. E. (2014). *Neuroanatomical Basis of Clinical Neurology*. CRC Press.

Ataloglou, D., Dimou, A., Zarpalas, D., and Daras, P. (2019). Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. *Neuroinformatics* 17, 563–582. doi: 10.1007/s12021-019-09417-y

Bartsch, T., Döhring, J., Rohr, A., Jansen, O., and Deuschl, G. (2011). Ca1 neurons in the human hippocampus are critical for autobiographical memory, mental time travel, and autonoetic consciousness. *Proc. Natl. Acad. Sci. U.S.A.* 108, 17562–17567. doi: 10.1073/pnas.1110266108

Beattie, J. F., Martin, R. C., Kana, R. K., Deshpande, H., Lee, S., Curé, J., et al. (2017). Hippocampal dentation: structural variation and its association with episodic memory in healthy adults. *Neuropsychologia* 101, 65–75. doi: 10.1016/j.neuropsychologia.2017.04.036

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

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

Bohbot, V. D., Iaria, G., and Petrides, M. (2004). Hippocampal function and spatial memory: evidence from functional neuroimaging in healthy participants and performance of patients with medial temporal lobe resections. *Neuropsychology* 18, 418. doi: 10.1037/0894-4105.18.3.418

Cates, J., Fletcher, P. T., Styner, M., Hazlett, H. C., and Whitaker, R. (2008). "Particle-based shape analysis of multi-object complexes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 477–485.

Chang, C., Huang, C., Zhou, N., Li, S. X., Ver Hoef, L., and Gao, Y. (2018). The bumps under the hippocampus. *Hum. Brain Mapp.* 39, 472–490. doi: 10.1002/hbm.23856

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

Convit, A., De Leon, M., Tarshish, C., De Santi, S., Tsui, W., Rusinek, H., et al. (1997). Specific hippocampal volume reductions in individuals at risk for Alzheimer's disease. *Neurobiol. Aging* 18, 131–138. doi: 10.1016/S0197-4580(97)00001-8

Derix, J., Yang, S., Lüsebrink, F., Fiederer, L. D. J., Schulze-Bonhage, A., Aertsen, A., et al. (2014). Visualization of the amygdalo–hippocampal border and its structural variability by 7T and 3T magnetic resonance imaging. *Hum. Brain Mapp.* 35, 4316–4329. doi: 10.1002/hbm.22477

Duvernoy, H. M. (2013). *The Human Hippocampus: An Atlas of Applied Anatomy*. JF Bergmann-Verlag.

Duvernoy, H. M., Cattin, F., and Risold, P.-Y. (2005). *The Human Hippocampus: Functional Anatomy, Vascularization and Serial Sections With MRI*. Springer.

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

Fleisher, A., Sun, S., Taylor, C., Ward, C., Gamst, A., Petersen, R., et al. (2008). Volumetric MRI vs clinical predictors of Alzheimer disease in mild cognitive impairment. *Neurology* 70, 191–199. doi: 10.1212/01.wnl.0000287091.57376.65

Fraser, M. A., Shaw, M. E., and Cherbuin, N. (2015). A systematic review and meta-analysis of longitudinal hippocampal atrophy in healthy human ageing. *Neuroimage* 112, 364–374. doi: 10.1016/j.neuroimage.2015.03.035

Frisoni, G. B., Jack, C. R. Jr., Bocchetta, M., Bauer, C., Frederiksen, K. S., Liu, Y., et al. (2015). The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. *Alzheimers Dement.* 11, 111–125. doi: 10.1016/j.jalz.2014.05.1756

Gao, Y., and Bouix, S. (2016). Statistical shape analysis using 3D Poisson equation–a quantitatively validated approach. *Med. Image Anal.* 30, 72–84. doi: 10.1016/j.media.2015.12.007

Gao, Y., Riklin-Raviv, T., and Bouix, S. (2014). Shape analysis, a field in need of careful validation. *Hum. Brain Mapp.* 35, 4965–4978. doi: 10.1002/hbm.22525

Gao, Y., and Tannenbaum, A. (2010). "Image processing and registration in a point set representation," in *Medical Imaging 2010: Image Processing* (SPIE), 84–92.

Gerig, G., Styner, M., Jones, D., Weinberger, D., and Lieberman, J. (2001). "Shape analysis of brain ventricles using spharm," in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)* (IEEE), 171–178.

Göritz, C., and Frisén, J. (2012). Neural stem cells and neurogenesis in the adult. *Cell Stem Cell* 10, 657–659. doi: 10.1016/j.stem.2012.04.005

Goubran, M., Ntiri, E. E., Akhavein, H., Holmes, M., Nestor, S., Ramirez, J., et al. (2020). *Hippocampal Segmentation for Brains With Extensive Atrophy Using Three-Dimensional Convolutional Neural Networks*. Technical report, Wiley Online Library.

Guo, Y., Wu, Z., and Shen, D. (2020). Learning longitudinal classification-regression model for infant hippocampus segmentation. *Neurocomputing* 391, 191–198. doi: 10.1016/j.neucom.2019.01.108

Henke, K., Weber, B., Kneifel, S., Wieser, H. G., and Buck, A. (1999). Human hippocampus associates information in memory. *Proc. Natl. Acad. Sci. U.S.A.* 96, 5884–5889. doi: 10.1073/pnas.96.10.5884

Hong, Y., Gao, Y., Niethammer, M., and Bouix, S. (2015). Shape analysis based on depth-ordering. *Med. Image Anal.* 25, 2–10. doi: 10.1016/j.media.2015.04.004

IXI dataset (2018). Available online at: https://brain-development.org/ixi-dataset/

Khachaturyan, A., Semenovsovskaya, S., and Vainshtein, B. (1981). The thermodynamic approach to the structure analysis of crystals. *Acta Crystallogr. A* 37, 742–754. doi: 10.1107/S0567739481001630

Kilpattu Ramaniharan, A., Zhang, M. W., Selladurai, G., Martin, R., and Ver Hoef, L. (2022). Loss of hippocampal dentation in hippocampal sclerosis and its relationship to memory dysfunction. *Epilepsia* 63, 1104–1114. doi: 10.1111/epi.17211

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

Knoth, R., Singec, I., Ditter, M., Pantazis, G., Capetian, P., Meyer, R. P., et al. (2010). Murine features of neurogenesis in the human hippocampus across the lifespan from 0 to 100 years. *PLoS ONE* 5, e8809. doi: 10.1371/journal.pone.0008809

Konishi, K., Mckenzie, S., Etchamendy, N., Roy, S., and Bohbot, V. D. (2017). Hippocampus-dependent spatial learning is associated with higher global cognition among healthy older adults. *Neuropsychologia* 106, 310–321. doi: 10.1016/j.neuropsychologia.2017.09.025

Kraguljac, N. V., White, D. M., Reid, M. A., and Lahti, A. C. (2013). Increased hippocampal glutamate and volumetric deficits in unmedicated patients with schizophrenia. *JAMA Psychiatry* 70, 1294–1302. doi: 10.1001/jamapsychiatry.2013.2437

Lim, D. A., and Alvarez-Buylla, A. (2016). The adult ventricular–subventricular zone (V-SVZ) and olfactory bulb (OB) neurogenesis. *Cold Spring Harbor Perspect. Biol.* 8, a018820. doi: 10.1101/cshperspect.a018820

Liu, M., Li, F., Yan, H., Wang, K., Ma, Y., Shen, L., et al. (2020). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage* 208, 116459. doi: 10.1016/j.neuroimage.2019.116459

Luders, E., Narr, K. L., Bilder, R. M., Szeszko, P. R., Gurbani, M. N., Hamilton, L., et al. (2008). Mapping the relationship between cortical convolution and intelligence: effects of gender. *Cereb. Cortex* 18, 2019–2026. doi: 10.1093/cercor/bhm227

Memmel, M., Gonzalez, C., and Mukhopadhyay, A. (2021). "Adversarial continual learning for multi-domain hippocampal segmentation," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health* (Springer), 35–45.

Nestor, S. M., Gibson, E., Gao, F.-Q., Kiss, A., Black, S. E., and Alzheimer's Disease Neuroimaging Initiative (2013). A direct morphometric comparison of five labeling protocols for multi-atlas driven automatic segmentation of the hippocampus in alzheimer's disease. *Neuroimage* 66, 50–70. doi: 10.1016/j.neuroimage.2012.10.081

Pang, S., Lu, Z., Jiang, J., Zhao, L., Lin, L., Li, X., et al. (2019). Hippocampus segmentation based on iterative local linear mapping with representative and local structure-preserved feature embedding. *IEEE Trans. Med. Imaging* 38, 2271–2280. doi: 10.1109/TMI.2019.2906727

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* 56, 907–922. doi: 10.1016/j.neuroimage.2011.02.046

Ribas, G. C. (2018). *The Cerebral Architecture*. Cambridge University Press.

Riklin Raviv, T., Gao, Y., Levitt, J. J., and Bouix, S. (2014). Statistical shape analysis of neuroanatomical structures via level-set–based shape morphing. *SIAM J. Imaging Sci.* 7, 1645–1668. doi: 10.1137/13093978X

Sanai, N., Nguyen, T., Ihrie, R. A., Mirzadeh, Z., Tsai, H.-H., Wong, M., et al. (2011). Corridors of migrating neurons in the human brain and their decline during infancy. *Nature* 478, 382–386. doi: 10.1038/nature10487

Scher, A. I., Xu, Y., Korf, E., White, L. R., Scheltens, P., Toga, A. W., et al. (2007). Hippocampal shape analysis in Alzheimer's disease: a population-based study. *Neuroimage* 36, 8–18. doi: 10.1016/j.neuroimage.2006.12.036

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L., Trojanowski, J., et al. (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to apoe genotype and biomarkers. *Brain* 132, 1067–1077. doi: 10.1093/brain/awp007

Shen, L. (2010). *SPHARM-MAT v1. 0.0 Documentation*.

Shen, L., Farid, H., and McPeek, M. A. (2009). Modeling three-dimensional morphological structures using spherical harmonics. *Evol. Int. J. Organ. Evol.* 63, 1003–1016. doi: 10.1111/j.1558-5646.2008.00557.x

Shen, L., and Makedon, F. (2006). Spherical mapping for processing of 3D closed surfaces. *Image Vision Comput.* 24, 743–761. doi: 10.1016/j.imavis.2006.01.011

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). "SegmentER: transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7262–7272.

Styner, M., Lieberman, J. A., Pantazis, D., and Gerig, G. (2004). Boundary and medial shape analysis of the hippocampus in schizophrenia. *Med. Image Anal.* 8, 197–203. doi: 10.1016/j.media.2004.06.004

Styner, M., Oguz, I., Xu, S., Brechbühler, C., Pantazis, D., Levitt, J. J., et al. (2006). Framework for the statistical shape analysis of brain structures using SPHARM-PDM. *Insight J.* 242–250. doi: 10.54294/owxzil

ten Donkelaar, H. J., Kachlík, D., and Tubbs, R. S. (2018). *An Illustrated Terminologia Neuroanatomica: A Concise Encyclopedia of Human Neuroanatomy*. Springer.

ten Hove, J., and Poppenk, J. (2020). Structural variation in hippocampal dentations among healthy young adults. *bioRxiv*. doi: 10.1101/2020.02.09.940726

Thompson, P. M., Hayashi, K. M., De Zubicaray, G. I., Janke, A. L., Rose, S. E., Semple, J., et al. (2004). Mapping hippocampal and ventricular change in Alzheimer disease. *Neuroimage* 22, 1754–1766. doi: 10.1016/j.neuroimage.2004.03.040

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

Tian, M., He, J., Yu, X., Cai, C., and Gao, Y. (2021). MCMC guided CNN training and segmentation for pancreas extraction. *IEEE Access* 9, 90539–90554. doi: 10.1109/ACCESS.2021.3070391

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., and Chandraker, M. (2018). "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7472–7481.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021). "Medical transformer: gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 36–46.

Van Opbroek, A., Achterberg, H. C., Vernooij, M. W., and De Bruijne, M. (2018). Transfer learning for image segmentation by combining image weighting and kernel learning. *IEEE Trans. Med. Imaging* 38, 213–224. doi: 10.1109/TMI.2018.2859478

Walker, T. L., Turnbull, G. W., Mackay, E. W., Hannan, A. J., and Bartlett, P. F. (2011). The latent stem cell population is retained in the hippocampus of transgenic Huntington's disease mice but not wild-type mice. *PLoS ONE* 6, e18153. doi: 10.1371/journal.pone.0018153

Wang, L., Miller, J. P., Gado, M. H., McKeel, D. W., Rothermich, M., Miller, M. I., et al. (2006). Abnormalities of hippocampal surface structure in very mild dementia of the Alzheimer type. *Neuroimage* 30, 52–60. doi: 10.1016/j.neuroimage.2005.09.017

Wisse, L., Gerritsen, L., Zwanenburg, J. J., Kuijf, H. J., Luijten, P. R., Biessels, G. J., et al. (2012). Subfields of the hippocampal formation at 7T MRI: *in vivo* volumetric assessment. *Neuroimage* 61, 1043–1049. doi: 10.1016/j.neuroimage.2012.03.023

Wu, K., Taki, Y., Sato, K., Qi, H., Kawashima, R., and Fukuda, H. (2013). A longitudinal study of structural brain network changes with normal aging. *Front. Hum. Neurosci.* 7, 113. doi: 10.3389/fnhum.2013.00113

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). "COTR: efficiently bridging CNN and transformer for 3D medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 171–180.

Yushkevich, P. A., Avants, B. B., Pluta, J., Das, S., Minkoff, D., Mechanic-Hamilton, D., et al. (2009). A high-resolution computational atlas of the human hippocampus from postmortem magnetic resonance imaging at 9.4 T. *Neuroimage* 44, 385–398. doi: 10.1016/j.neuroimage.2008.08.042

Zavaliangos-Petropulu, A., Tubi, M. A., Haddad, E., Zhu, A., Braskie, M. N., Jahanshad, N., et al. (2022). Testing a convolutional neural network-based hippocampal segmentation method in a stroke population. *Hum. Brain Mapp.* 43, 234–243. doi: 10.1002/hbm.25210

Zuccato, C., Valenza, M., and Cattaneo, E. (2010). Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiol. Rev.* 90, 905–981. doi: 10.1152/physrev.00041.2009

frontiers | Frontiers in Neuroscience

TYPE Methods
PUBLISHED 07 November 2023
DOI 10.3389/fnins.2023.1213099

OPEN ACCESS

EDITED BY
Ateke Goshvarpour,
Imam Reza International University, Iran

REVIEWED BY
Man Fai Leung,
Anglia Ruskin University, United Kingdom
Hongli Chang,
Southeast University, China
Jianxing Liu,
Harbin Institute of Technology, China

*CORRESPONDENCE
Di Zhou
✉ 1008488@qq.com;
✉ 2019025@nbpt.edu.cn

†These authors have contributed equally to this work

RECEIVED 27 April 2023
ACCEPTED 04 October 2023
PUBLISHED 07 November 2023

CITATION
Tao J, Dan Y and Zhou D (2023) Local domain generalization with low-rank constraint for EEG-based emotion recognition.
Front. Neurosci. 17:1213099.
doi: 10.3389/fnins.2023.1213099

Local domain generalization with low-rank constraint for EEG-based emotion recognition

Jianwen Tao[1†], Yufang Dan[1†] and Di Zhou[2*]

[1]Institute of Artificial Intelligence Application, Ningbo Polytechnic, Zhejiang, China, [2]Industrial Technological Institute of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, China

As an important branch in the field of affective computing, emotion recognition based on electroencephalography (EEG) faces a long-standing challenge due to individual diversities. To conquer this challenge, domain adaptation (DA) or domain generalization (i.e., DA without target domain in the training stage) techniques have been introduced into EEG-based emotion recognition to eliminate the distribution discrepancy between different subjects. The preceding DA or domain generalization (DG) methods mainly focus on aligning the global distribution shift between source and target domains, yet without considering the correlations between the subdomains within the source domain and the target domain of interest. Since the ignorance of the fine-grained distribution information in the source may still bind the DG expectation on EEG datasets with multimodal structures, multiple patches (or subdomains) should be reconstructed from the source domain, on which multi-classifiers could be learned collaboratively. It is expected that accurately aligning relevant subdomains by excavating multiple distribution patterns within the source domain could further boost the learning performance of DG/DA. Therefore, we propose in this work a novel DG method for EEG-based emotion recognition, i.e., Local Domain Generalization with low-rank constraint (LDG). Specifically, the source domain is firstly partitioned into multiple local domains, each of which contains only one positive sample and its positive neighbors and k_2 negative neighbors. Multiple subject-invariant classifiers on different subdomains are then co-learned in a unified framework by minimizing local regression loss with low-rank regularization for considering the shared knowledge among local domains. In the inference stage, the learned local classifiers are discriminatively selected according to their importance of adaptation. Extensive experiments are conducted on two benchmark databases (DEAP and SEED) under two cross-validation evaluation protocols, i.e., cross-subject within-dataset and cross-dataset within-session. The experimental results under the 5-fold cross-validation demonstrate the superiority of the proposed method compared with several state-of-the-art methods.

KEYWORDS

domain adaptation, subdomain generalization, emotion recognition, electroencephalogram, local learning

Introduction

In the field of affective computing research (Mühl et al., 2014), automatic emotion recognition (AER; Dolan, 2002) has received considerable attention from computer vision communities (Kim et al., 2013). Many EEG-based emotion recognition methods have been proposed so far (Musha et al., 1997; Jenke et al., 2014; Zheng, 2017; Niu et al., 2018; Pandey and Seeja, 2019; Chang et al., 2021, 2023; Zhou et al., 2022). From the viewpoint of machine learning, EEG-based AER can be modeled as a classification or regression problem (Kim et al., 2013; Zhang et al., 2017), in which state-of-the-arts for AER usually tailor their classifiers trained on multiple subjects and apply them to individual subjects. From both qualitative and empirical observations, the generalizability of AER could be attenuated partly due to the individual differences among subjects (Jayaram et al., 2016; Zheng and Lu, 2016; Lan et al., 2018). That is, the subject-independent classifier usually achieves an inferior generalization performance since emotion patterns may significantly vary from one subject to another (Pandey and Seeja, 2019). As a possible solution, subject-specific classifiers are usually impractical due to insufficient training data (Li X. et al., 2018; Zhou et al., 2022). While conspicuous progress has been made to conquer this issue by improving feature representations and learning models (Zheng and Lu, 2015; Song et al., 2018; Li et al., 2018a,b; Li Y. et al., 2019; Du et al., 2020; Zhong P. et al., 2020; Zhou et al., 2022), there still exists a long-standing challenge incurred by individual diversities in EEG-based AER. This challenge is primarily attributed to the fact that the learned classifiers should be generalized into previously unseen subjects that may obviously differ from those on which the classifiers are trained (Ghifary et al., 2017). To this end, numerous domain adaptation (DA) learning algorithms for AER have emerged by exploiting EEG features (Zheng et al., 2015; Chai et al., 2017; Li J. et al., 2019; Pandey and Seeja, 2019; Zhang et al., 2019b; Li et al., 2020; Chen et al., 2021; Dan et al., 2021; Tao et al., 2022). For instance, Pandey and Seeja (2019)) and Li X. et al. (2018) successively proposed two subject invariant models for EEG-based emotion recognition; following the deep network architecture, in the researchers (Chai et al., 2016; Li H. et al., 2018; Luo et al., 2018; Li et al., 2018c, 2021; Wang et al., 2022; Zhou et al., 2022) designed several deep learning models for EEG-based emotion recognition.

Unfortunately, in some practical AER applications, the whole target data of interest may be unavailable in the stage of training a subject-specific classifier (Wang et al., 2022). In this case, domain generalization (DG; Muandet et al., 2013), an effective variant of DA (Bruzzone and Marconcini, 2010), is proved to be a feasible solution for DA emotion recognition (Tao et al., 2022). With no need to focus on the generalization of some specific target domain, DG methodology could better acquire out-of-the-distribution effects on test samples from other previously unseen target domains (Wang et al., 2022). While DA and DG are closely related in learning scenarios, DA algorithms generally cannot be directly applicable to DG since they rely on the availability of the target domain in the stage of training. In this sense, DG is more challenging than DA as no target data can be used for fine-tuning in the training stage (Ghifary et al., 2017).

In DA/DG, one major problem is how to reduce or eliminate the distribution discrepancy between different domains (Patel et al., 2015; Wang et al., 2022). First of all, one needs to design a robust and effective criterion that can measure the domain discrepancy. Due to its simplicity, effectiveness, and intuition, Maximum Mean Discrepancy (MMD; Gretton et al., 2009) is a commonly adopted distribution distance measure criterion. Preceding MMD-based DA methods (Pan et al., 2011; Duan et al., 2012; Tao et al., 2012, 2017, 2019; Chen et al., 2013; Long et al., 2014a; Ding et al., 2018a,b,c), however, generally focused on the global statistical distribution shift between/among different domains without considering the complementarities and diversities between two subdomains constructed with local structures within the same/different domains (Gao et al., 2015; Zhu et al., 2020). This could result in attenuated adaptation performance to some extent, since not only could all the samples from both source and target domains be confused together, but also the local discriminative structures could be trimmed without capturing the fine-grained local structures (Zhu et al., 2020). That is, while the global distribution alignment may lead to approximate zero distribution distance between different domains, a common challenge that exists in preceding global methods is that the samples from different domains are pulled too close to be accurately classified. An intuitive example is shown in Figure 1, where the source domain presents a certain multimodal structure (as shown in Figure 1A). After global domain adaptation, as shown in Figure 1B, the distributions of the two domains are approximately the same, but the data in different semantic structures are too close to be classified accurately. This is a common problem in previous global DA methods. Hence, matching the global source and target domains may not work well in this scenario.

Concerning the challenge of global domain shift, several works pay attention to semantic alignment or matching conditional distribution (Long et al., 2014a, 2017). There are other works proposed to discover multiple latent domains by decomposing the source domain (Judy et al., 2012; Gao et al., 2015). While they have presented the effectiveness of DA by exploring multiple subdomains potentially existing in the source domain, discovering multiple representative latent domains is still a non-trivial task by explicitly dividing the source samples into multiple blobs (Zhu et al., 2020). Further, to overcome the shortages that exist in the global distribution measure, numerous deep subdomain adaptation methods have focused on accurately aligning the distributions between different subdomains (Gao et al., 2015; Zhu et al., 2020). For instance, the recent work in Zhu et al. (2020) focuses on aligning the distribution of the relevant subdomains within the same category in the source and target domains. These deep learning methods, however, usually contain several updatable loss functions and converge slowly. Moreover, it is still an unexplained open problem whether the success of deep DA methods really benefits from the feature representations, fine-tuned classifiers, or effects of the adaptation regularizers (Tao et al., 2022).

Motivated by the idea of subdomain adaptation, we propose in this work a Local Domain Generalization (LDG) scheme to implicitly align the relevant local domain distributions from a single source with that of the target domain. A key improvement of LDG over previous DG/DA methods is the capability of the fine-grained alignment of a domain shift by capturing the local discriminative structures in the source domain by excavating multiple subdomains as per each positive sample with its two k-NN subsets (as shown in Figure 1C). In these local domains, multiple classifiers can be jointly trained in a unified framework by aligning them with a referenced model. Under this framework, the model discrepancies between the relevant subdomains from the source and the target domain could be measured by

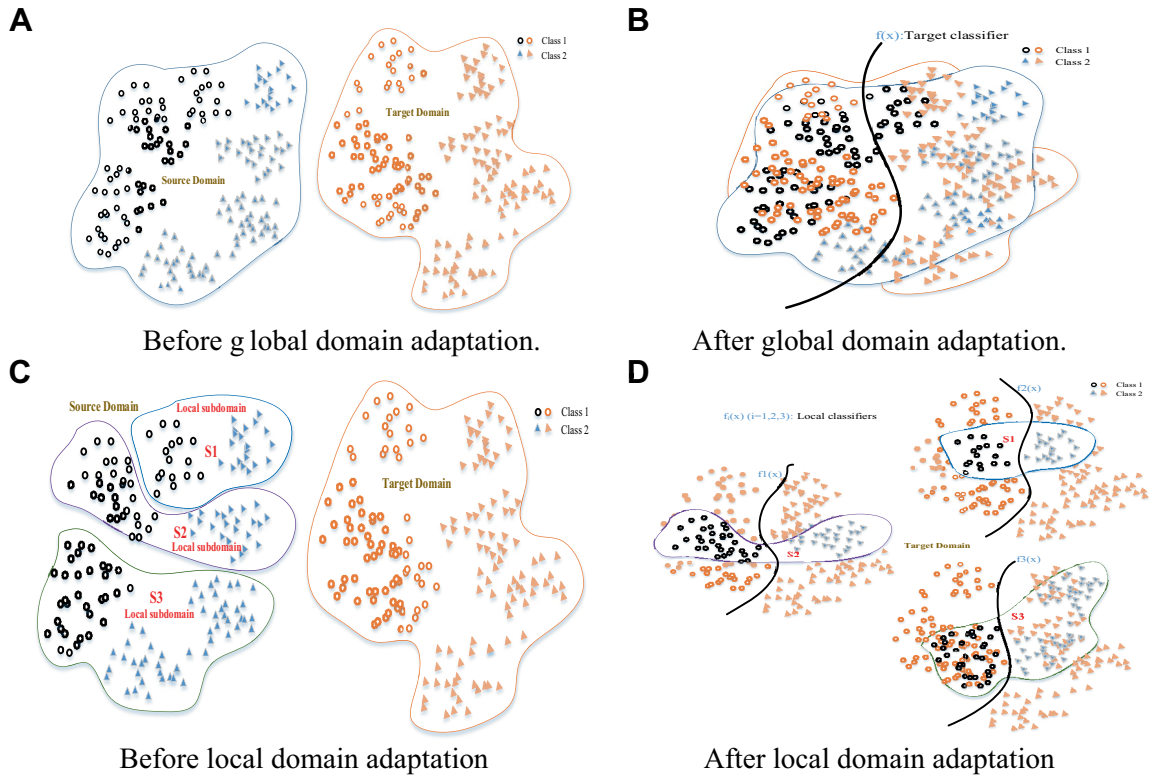


FIGURE 1
Global domain adaptation might lose some fine-grained information **(A,B)**. Local domain adaptation can exploit the local discriminative structures to capture the fine-grained information for each category **(C,D)**.

considering the weights as per different distribution distances. After local domain adaptation, as shown in Figure 1D, each local domain distribution from the source domain is approximately the same as that of the target domain. Therefore, multiple local classifiers jointly learned with these local domain adaptations could be integrated and generalized into the target domain.

Specifically, we present an LDG framework for AER with EEG features with low-rank constraints. Under this framework, the source domain is firstly divided into multiple local domains, each containing only one positive sample (or exemplar; Zhang et al., 2016) and its positive and k_2 negative neighbors. Intuitively, the distribution structures of these local domains for those exemplars are expected to be relatively closer and simpler than that of the global one. In LDG, multiple subject-invariant classifiers on different local domains are co-learned in a joint framework by minimizing local regression loss. Instead of evaluating the importance of each classifier individually, LDG selects models in a collaborated mode by considering the shared knowledge among local domains by additionally imposing a nuclear-norm-based regularizer on the objective function. The learned local classifiers are discriminatively selected according to their weights in the inference stage. While the DG performance of LDG also can be boosted with most feedforward network models by exploiting the deep feature representations, it does not need iterative deep training and converges fast, thus being very efficient and effective.

Different from the existing DG methods that only focus on global distribution alignment in the source domain(s), we consider the local distribution structures of the source domain and their relevance with the target domain to further enhance the effectiveness and generalizability of the learned adaptation model. Our algorithm can adapt as much knowledge as possible from a certain source domain, even if the EEG features between domains are partially distinct but overlapping. To the best of our knowledge, there is no prior work imposing DG with multiple local domains on solving AER problems. The main contributions of this paper are summarized as follows.

1. We propose a local domain generalization framework (LDG) for EEG-based emotion recognition by leveraging multiple structure-similar local domains from the source domain with multi-model distribution patterns. Using this framework, the capacity of MMD-based DA methods can be extended by excavating the local discriminative structures for each domain by aligning KNN-based local domain distributions.
2. We present a subdomain division strategy, i.e., splitting the source domain into multiple local domains, each of which is composed of each positive (exemplar) sample (Zhang et al., 2016; Li W. et al., 2018; Niu et al., 2018) and its k_1 positive and k_2 negative neighbors. Multiple local classifiers can be, respectively, trained on each local domain. We then formulate a new objective function by imposing a nuclear-norm-based regularizer on the model matrix in the objective function to further enhance the discriminative capability of the learned local classifiers by exploiting the intrinsic discriminative structure in the source domain.

3. An iterative optimization algorithm is presented for solving the objective of LDG that can be applied to EEG-based AER problems. The convergence of the optimization procedure can be guaranteed in terms of the proof of the proposed convergence theorem.
4. Extensive experiments are conducted on two benchmark databases (DEAP and SEED) under two cross-validation evaluation protocols (cross-subject within-dataset and cross-dataset within-session). The remarkable experimental results show that our method outperforms other state-of-the-art methods on emotion recognition tasks.

The rest of the paper is organized as follows. Section 2 reviews several related works in emotion recognition, DG, and subdomain adaptation. Section 3 introduces our LDG framework including the overall objective function, and then the optimization algorithm and its convergence analysis are successively provided in Section 4. Section 5 provides a series of experiments to evaluate the effectiveness of LDG for AER. Finally, we summarize the entire paper in Section 6.

Related work

In recent decades, increasing attention has been given to emotion recognition with brain-computer interfaces (BCI; Dolan, 2002; Kim et al., 2013; Mühl et al., 2014) in the affective computing community. A vanilla aBCI system using spontaneous EEG signals firstly extracts sufficient discriminative features from the EEG data by a certain feature extractor and then trains an optimal classifier using these features and the corresponding emotion states for AER. Therefore, a proper design of EEG-based emotion recognition models helps facilitate the data processing, benefits from discriminant feature characterization, and lightens the model performance. The latest works about affective BCI (aBCI) usually adopt machine learning algorithms on automatic emotion recognition (AER) using extracted discriminative features (Musha et al., 1997; Jenke et al., 2014; Chang et al., 2023). However, the traditional machine learning method has a major disadvantage in that the feature extraction process is usually cumbersome, and relies heavily on human experts. Then, end-to-end deep learning methods emerged as an effective way to address this disadvantage with the help of raw EEG signals and time-frequency spectrums (Han et al., 2022). More details can be found in Zhang et al. (2020c), which investigated the application of several deep learning models to the research field of EEG-based emotion recognition, including deep neural networks (DNN) (Chang et al., 2021), convolutional neural networks (CNN), long short-term memory (LSTM), and a hybrid model of CNN and LSTM (CNN-LSTM; Zhong Q. et al., 2020; Mughal et al., 2022; Xu et al., 2022).

While preceding methods have obtained remarkable achievements on EEG-based AER (Zheng, 2017; Li et al., 2018a,b; Li Y. et al., 2019; Pandey and Seeja, 2019), the performance expectation for cross-subject/dataset recognition could be lowered due to the diversities of emotional states among subjects/datasets (Jayaram et al., 2016; Zheng and Lu, 2016; Li X. et al., 2018). While subject-specific classifiers may be a possible solution for these cases, they are usually infeasible in real tasks due to insufficient training data. Moreover, even if they are feasible in some specific scenarios, it is also an indispensable task to fine-tune the classifier to maintain a sound recognition capacity partly because the EEG signals of the same subject sometimes change (Zhou et al., 2022). Fortunately, the recently proposed domain adaptation (DA) technique (Patel et al., 2015) can be leveraged to surmount these challenges for EEG-based emotion recognition. As a well-focused research direction, the unsupervised domain adaptation (UDA) methodology has promoted a large amount of research effort devoted to generalizing the knowledge learned from one/multiple labeled source domain(s) into a different but related unlabeled target domain (Wang and Mahadevan, 2011; Gong et al., 2012; Long et al., 2014b, 2015, 2016; Ganin and Lempitsky, 2015; Ganin et al., 2016; Judy et al., 2017; Tzeng et al., 2017; Ding et al., 2018a,b,c). Over the past decade, DA-based emotion recognition methods have been a hot topic (Lan et al., 2018), almost fully covered in the literature of aBCI (Zheng et al., 2015; Chai et al., 2016, 2017; Jayaram et al., 2016; Zheng and Lu, 2016; Li H. et al., 2018; Li X. et al., 2018; Luo et al., 2018; Li et al., 2018c, 2020, 2021; Li J. et al., 2019; Chen et al., 2021; Dan et al., 2021; Tao et al., 2022; Zhou et al., 2022). Existing methods explore tackling different challenges in AER with EEG datasets by excavating a certain latent subspace shared by different domains for filling the domain distance among subjects or sessions.

In some real DA-based AER applications, the whole target data of interest may be unavailable in the stage of training (Ghifary et al., 2017). In this scenario, domain generalization (DG; Muandet et al., 2013), an effective variant of DA, has been proven to be a feasible solution for DA emotion recognition since it need not focus on the generalization of a certain specific target domain. While DA and DG are closely related in learning scenarios, DA algorithms generally are not directly applicable to DG since they rely on the availability of the target domain in the stage of training. In this sense, DG is more challenging than DA as no target data can be used for fine-tuning in the training stage. The extant works about DG can be divided into two research lines in terms of different strategies, i.e., feature-centric DG (Judy et al., 2012; Muandet et al., 2013; Ghifary et al., 2017; Motiian et al., 2017) and classifier-centric DG (Xu et al., 2014; Ghifary et al., 2015; Niu et al., 2015, 2018; Gan et al., 2016; Li W. et al., 2018). The former aims to mine domain-invariant features, while the latter uses multi-classifiers adaptation by regulating their weights. More research progress on DG can be found in the recent survey on DG (Wang et al., 2022).

As is known, a major task in vanilla UDA/DG methodology is to mitigate the domain discrepancy either by aligning the statistical moments (Pan et al., 2011; Duan et al., 2012; Tao et al., 2012; Chen et al., 2013; Long et al., 2014a,b; Xiao and Guo, 2015; Ding et al., 2018a,b,c) or by using domain adversarial learning (Ganin and Lempitsky, 2015; Ganin et al., 2016; Tzeng et al., 2017; Long et al., 2018; Pei et al., 2018) benefited from the powerful deep neural networks. Traditional DA/DG methods usually assume a global distribution shift between different domains and expect approximately the same global distribution of two domains after adaptation (Mansour et al., 2009). However, most of the preceding DA/DG methods face a common problem in that they only pay attention to matching the global statistical distribution between domains without considering the complementarities and diversities among subdomains constructed using several local structures within the same/different domains (Zhu et al., 2020). This could result in attenuated adaptation performance in part because the samples from

different domains are pulled too close to be accurately classified in those global methods. As a result, not only will all the data from the source and target domains be confused, but also the discriminative structures can be mixed up. Subdomain adaptation can to some extent conquer the shortcomings in aligning global domain discrepancy. For instance, several related works have been proposed to excavate multiple latent domains from the source domain (Judy et al., 2012). To discover multiple representative latent domains, however, is a non-trivial task done by explicitly dividing the source samples into multiple blobs. Aiming at the disadvantages of global domain adaptation, considerable works (Gao et al., 2015; Zhu et al., 2020) have explored subdomain adaptation, which focuses on aligning the local domain discrepancies. Most deep DA/DG methods belong to the deep adversarial learning methodology and converge slowly due to several loss functions. To this end, Zhu et al. (2020) recently presented a deep subdomain adaptation network (DSAN) based on the proposed local maximum mean discrepancy (LMMD), which learns a DA network by aligning the related distributions of subdomains across different domains.

It is worth noting that the discriminative structures could still be mixed up in extant subdomain adaptation schemes when the source (or target) domain presents a multimodal distribution structure (as shown in Figure 1). Different from these works on aligning global/sub-domain(s) shift(s), we propose a novel fine-grained DG method for EEG-based emotion recognition, in which multiple patches (local domains) are firstly reconstructed from the source dataset and multiple local classifiers are then learned collaboratively for effective generalization into the target domain even with multiple kinds of distribution pattern (Gao et al., 2015). Our method does not need deep training and converges fast, while its adaptation expectation can be easily boosted with deep feature representations from most feedforward network models.

Proposed framework

Preliminary notations

In the context of this paper, we, respectively, denote by small and capital letters the column vectors and matrices. The frequently used notations are summarized in Table 1. The concatenation operations of matrices along the row (horizontally) are denoted as $[A_1, A_2, \ldots, A_k]$, and their concatenation along the column (vertically) are denoted as $[A_1, A_2, \ldots, A_k]$.

Definition 1 (Local domain): For a certain domain $X = \{x_i\}_{i=1}^{m}$ with some probability distribution P, a local domain for one positive example $x_v \in X$ is composed of its k_1 positive nearest neighbor set $N_{k_1}^{+}(x_v) = \{x_{v_1}, \ldots, x_{v_{k_1}}\}$ and k_2 negative neighbor set $N_{k_2}^{-}(x_v) = \{x_{v_{k_1+1}}, \ldots, x_{v_{k_1+k_2}}\}$, i.e., $X_v = \{x_v, N_{k_1}^{+}(x_v), N_{k_2}^{-}(x_v)\}$.

According to Definition 1, for any source domain $X^s = \{x_i^s\}_{i=1}^{n_s}$ with p positive samples $\{x_v^s \in \mathbb{R}^d\}_{v=1}^{p}$ and $n_s - p$ negative samples, one can reconstruct p local domains $X_v^s = \{x_v^s, N_{k_1}^{+}(x_v^s), N_{k_2}^{-}(x_v^s)\}$, $1 \le v \le p$, by finding the positive nearest neighbor set $N_{k_1}^{+}(x_v^s) = \{x_{v_1}^s, \ldots, x_{v_{k_1}}^s\}$ and k_2 negative neighbor set for each positive sample x_v^s $(1 \le v \le p)$.

TABLE 1 Notations and descriptions.

Notations	Descriptions
n	Data size.
d	Feature dimensionality of data.
χ	Data space.
Γ	Label space.
$a = [a_1, a_2, \ldots, a_d]^T \in \mathbb{R}^d$	Feature vector.
$A \in \mathbb{R}^{n \times d}$	Data matrix.
$A_{i,j}$	The (i, j) entry of A.
A_i and A_j	The i-th row and j-th column of A.
A^T and a^T	The transpose of matrix A and vector a.
$tr(A)$	The trace of a matrix A.
$\langle A_1, A_2 \rangle = tr\left(A_1^T A_2\right)$	The inner product of two matrices A_1 and A_2.
$\|a\|_p = \left(\sum_{i=1}^{d} \|a_i\|^p\right)^{1/p}$	The p-norm of a vector a.
$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{d} A_{i,j}^2}$	The Frobenius norm of A.
I_r	Identity matrix of size $r \times r$.
$\mathbf{1}_d$	d-dimensional vector of ones.
$\mathbf{0}_d$	d-dimensional vector of zeroes.

Definition 2 (Local domain adaptation, LDA): Let $\Delta = \{X_1^s, \ldots, X_m^s\}$ be a set of m local domains and $X^t \notin \Delta$ be a target domain. The task of LDA is to learn an ensemble function $f_{X^t}: \mathcal{X} \to \Gamma$ by co-learning multiple classifiers $f_v(X_v^s)$ $(1 \le v \le m)$ given Δ and X^t as the training examples by alleviating the distribution difference between source and target domains.

Definition 3 (Local domain generalization, LDG): In this scenario, the target domain is inaccessible in the training stage. Given m local domains $\Delta = \{X_1^s, \ldots, X_m^s\}$, and denoted by $X_a^s = \{x_i^a, y_i^a\}_{i=1}^{n_a}$ the samples drawn from the a-th subdomain, the task of LDG is to co-learn multiple adaptive functions $f_{X_a^s}: \mathcal{X} \to \Gamma$ only given $X_a^s, \forall a = 1, \ldots, m$ as the training examples, which could be well-generalized to a certain unseen target domain.

Motivation

As is known, a major task in vanilla UDA/DG methodology is to diminish the domain discrepancy either by aligning the statistical moments (Koelstra et al., 2012; Gao et al., 2015; Li et al., 2018a, 2020) or by domain adversarial learning (Gong et al., 2012; Lan et al., 2018; Li X. et al., 2018; Ding et al., 2018a) benefited from the powerful deep neural networks (Zhu et al., 2020; Zhou et al., 2022). While extensive exploration of cross-subject/session has been conducted effectively in

the prior works by leveraging various domain adaptation tricks, one obvious shortage in these works is they usually assume a global distribution shift between different subjects and expect an approximately similar global distribution of two subjects after adaptation. In other words, these DA-based AER methods only focus on matching the global statistical distribution between subjects without considering the complementarities and diversities among local domains constructed using some intrinsic structures within the same/different subjects. This leads to attenuated adaptation performance since the real-world EEG data is usually quite diverse and the distribution of emotion data is complex. It is challenging to reduce the global distribution discrepancy between different domains.

As far as we know, limited effort, however, has been witnessed in improving DA/DG performance by leveraging local knowledge among multiple subdomains from a single source. The ignorance of the fine-grained local discriminative structures may result in unsatisfying generalization capacity in DA/DG. Exploiting the relationships among multiple local domains to match their distribution divergences could not only align the global statistical distributions but also the local discriminative patterns. In many real applications, the local structure is more important than the global structure (Ding et al., 2018a), and the local learning algorithms often outperform global learning algorithms (Ding et al., 2018b). Because of this, LDA/LDG is able to compensate for the limitation of global DA since the diversities of domain distributions intrinsically exist in real applications.

Motivated by this idea, we propose in this paper a novel domain generalization framework for EEG-based emotion recognition, i.e., Local Domain Generalization (LDG) with low-rank constraints. Under this framework, LDA is a relaxed extension of LDG, where the target domain of interest is provided during the training process. Specifically, the source domain of the auxiliary is firstly partitioned into multiple local domains, each of which contains only one positive sample (or called exemplar sample) and its k_1 positive neighbors and k_2 negative neighbors. Each local domain is expected to be relatively more similar and possess a simpler distribution structure. Then multiple subject-invariant local classifiers are co-learned on these local domains by minimizing a unified local regression loss. Instead of evaluating the importance of each classification model individually, LDG selects models in a collaborated mode for considering the shared knowledge among local domains by additionally introducing a nuclear-norm-based regularizer into the objective function. In the inference stage, the learned local classifiers are discriminatively selected and reweighted according to the distribution distance between each local domain and the target domain of interest.

In the following sections, we will present the objective formulation of our framework followed by its effective optimization algorithm.

General formulation

In LDA/LDG learning, however, there still exists two challenges worthy to be effectively addressed: (1) how to divide one source into multiple local domains and (2) how to compute the weight of each sample in its local domain. Until now, little research has been reported to address these challenges for EEG-based emotion recognition through local regression learning by decomposing the source domain into multiple local domains. To address these challenges, in this section, we propose the general formulation of our framework underpinned by the robust local regression principle and the regularization theory. Concretely, our proposed method will possess several complementary characters, which can be combined into one unified optimization formulation so that a more effective target learning model and distribution alignment between local domains and the target domain can be jointly achieved.

For LDA of m local domains $\left\{X_v^s\right\}_{v=1}^m$ from the source domain X^s, we define the v-th ($1 \le v \le m$) local classifier as $f_v\left(w_v, X_v^s\right)$ corresponding to the v-th local domain, where $w_v \in \mathbb{R}^d$ is the v-th local classifier model. If we consider kernel learning and assume that there is a feature map $\phi_v : \chi \to H_v$[1] that projects the training data from the original feature space into a certain reproducing kernel Hilbert space (RKHS; Gretton et al., 2009) H_v, then the predictor model w_v can be kernelized. We denote the kernel matrix as $\left(K_v\right)_{i,j} = \phi\left(x_i^v\right), \phi\left(x_j^v\right)$, where $x_i^v, x_j^v \in X_v^s$. We introduce the empirical kernel map as discussed in Pan et al. (2011):

$$\begin{aligned}&\phi_v : \chi \to \mathbb{R}^d, \textit{ for linear kernel mapping}\\ &x \to K_v\left(\cdot, x^v\right)\Big|_{x_1^v, x_2^v, \ldots, x_{n_s}^v} = \left(K_v\left(x_1^v, x^v\right), \ldots, K_v\left(x_{n_s}^v, x^v\right)\right),\\ &\textit{for nonlinear kernel mapping}\end{aligned}$$

We therefore have kernelized data matrices $K_v^s = \phi_v\left(X_v^s\right)$ for nonlinear projection. For simplicity of expression, we uniformly express the data in linear and nonlinear space as follows:

$$\bar{X}_v^s = \begin{cases} X_v^s, \textit{ linear} \\ K_v^s\left(\cdot, x\right), \textit{ kernel} \end{cases}$$

In the sequence, we also refer to it as X_v^s (linear) and K_v^s (nonlinear) if without special denotation. We further denote by $W = \left[w_1;;\ldots;;w_m\right]$ the concatenated local model matrix. We then endeavor to find m local adaptation models parameterized by jointly exploiting correlation information among local domains.

We first formulate our method with classical regularized empirical error (Zhang et al., 2019c), which leads to a classifier f_v based on a set of training data X_v:

$$\min \sum_{v=1}^{m} loss\left(f_v\left(w_v, X_v\right), y_v\right) + \Omega\left(f_v\right) \qquad (1)$$

where $\Omega\left(f_v\right)$ is a regularization term that guarantees good generalization performance and $loss\left(\cdot,\cdot\right)$ is a regression loss function. Although other complex nonlinear models can be used, the linear model has the following characteristics: (1) It is fast and more suitable for practical applications and (2) The local structure of the manifold is approximately linear (Feiping Nie et al., 2010). So, we use the following linear transformation:

1 It is worthy to note that the feature mapping function ϕ_v ($1 \le v \le m$) with respect to each local domain can be completely different from each other.

$$f_v(w_v, X_v) = X_v^T w_v + b_v \tag{2}$$

where, $b_v \in \mathbb{R}$ is the bias term. The model vectors for all local domains should be highly correlated. So, we further get the following objective function.

$$\begin{aligned} &\min_{\theta_v, w_v, b_v} \sum_{v=1}^{m} \left\{ \theta_v{}^r \left\| X_v{}^T w_v + b_v 1_{k_1+k_2+1} - y_v \right\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \beta \|W\|_* \\ &s.t. \sum_{v=1}^{m} \theta_v = 1, \theta_v \in [0,1] \end{aligned} \tag{3}$$

where α, β is the regularization parameters and the coefficient θ_v is the contribution of each local model. The third term in Eq. (3) is the trace norm of $W \in \mathbb{R}^{d \times m}$, which is the convex hull of the rank of W, thus enhancing the correlation between different local weight vectors (Yang et al., 2013).

Essentially, it is expected that a bridge needs to be established between different local model vectors. Therefore, we can add a global model vector w and require each local model vector to be aligned with it (Zhang et al., 2019a). Furthermore, to avoid some noise information, we replace the real label vector y_v in Eq. (3) with the pseudo label vector $f_v \in \mathbb{R}^k$. This pseudo-label vector can be obtained by the subsequent optimization. Therefore, the objective function can be represented in the following formulation:

$$\begin{aligned} &\min_{\substack{\theta_v, w_v, b_v, \\ b, w}} \sum_{v=1}^{n_v} \left\{ \theta_v^r \left\| X_v^T w_v + b_v 1_k - f_v \right\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \sum_{v=1}^{n_v} \left\| X_v^T w_v - X_v^T \tilde{w} \right\|_2^2 \\ &+ \left\| X^T \tilde{w} + b 1_n - f \right\|_2^2 + \|f - y\|_2^2 + \beta \left(\|W\|_* + \|\tilde{w}\|_2^2 \right) \\ &s.t. \sum_{v=1}^{n_v} \theta_v = 1, \theta_v \in [0,1] \end{aligned} \tag{4}$$

where η is another regularization parameter. The reason for adding the fifth term is that the predicted results should be consistent with the actual label (Zhang et al., 2020a). We also expect that the local prediction label should be globally consistent, which is obtained by the global weight vector $\tilde{w}$ on each local domain. In other words, the label information should be consistent with the nearby samples.

Given our objectives mentioned above, we propose the following general formulation of LDG:

$$\begin{aligned} &\min_{\substack{\theta_v, w_v, f_v, f, \\ b_v, b, \tilde{w}, \lambda_v}} \sum_{v=1}^{m} \left\{ \theta_v^r \left\| X_v^T w_v + b_v 1_k - f_v \right\|_2^2 + \alpha \|w_v\|_2^2 \right\} + \sum_{v=1}^{m} \left\| X_v^T \mathbf{w}_v - X_v^T \tilde{w} \right\|_2^2 \\ &+ tr\left(\tilde{w}^T \left(\sum_{v=1}^{m} \lambda_v X_v L_v X_v^T \right) \tilde{w} \right) \\ &+ \left\| X^T \tilde{w} + b 1_n - f \right\|_2^2 + \|f - y\|_2^2 + \beta \left(\|W\|_* + \|\tilde{w}\|_2^2 \right) + \mu \sum_{v=1}^{m} \lambda_v \log \lambda_v \\ &s.t. \sum_{v=1}^{m} \theta_v = 1, \theta_v \in [0,1], \sum_{v=1}^{m} \lambda_v = 1, \lambda_v \in [0,1] \end{aligned} \tag{5}$$

where λ_v is the contribution of different subdomains. In the above equation, the maximum entropy regularization $\lambda_v \log \lambda_v$ is added to avoid a trivial solution. $L_v = (E_v)^{-1/2}(E_v - S_v)(E_v)^{-1/2}$ is a normalized Laplacian matrix corresponding to the v-th local domain (Yan et al., 2006), and E_v is a diagonal matrix with a diagonal element of $(E_v)_{i,i} = \sum_j^k (S_v)_{i,j}$. The graph weight matrix S_v of X_v is defined as follows:

$$(S_v)_{i,j} = \begin{cases} \exp\left(-\dfrac{x_i^v - x_j^{v2}}{\sigma^v} \right), & x_i^v \in \mathcal{N}_k(x_j^v) \text{or} x_j^v \in \mathcal{N}_k(x_i^v), \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{N}_k(x)$ denotes the k-NN of x.

Remark

In our objective formulation, one could adapt the knowledge obtained from multiple local domains to facilitate the target task of interest, which has been empirically demonstrated to be better than learning each local domain task independently in emotion recognition. In other words, it is expected to be beneficial to leverage the common knowledge shared by multiple local domain tasks for AER. However, most of the existing state-of-the-art algorithms uncover some optimal classifier models for the source and/or target domain independently. Moreover, in these state-of-the-art methods, joint multiple local adaptation learning has been largely unaddressed, and little or limited efforts have yet been devoted to the utilization of the correlation information among multiple local domains.

Optimization

Our objective function is non-smooth, so we propose an alternative algorithm to solve it.

Optimize b_v, w_v, f_v, f, b and $\tilde{w}$ by fixing λ_v, θ_v.

By setting the b_v derivative to 0, we have:

$$\begin{aligned} &w_v^T X_v \mathbf{1}_k + k b_v - f_v^T \mathbf{1}_k = 0 \\ &\Rightarrow b_v = \frac{1}{k}\left(f_v^T \mathbf{1}_k - w_v^T X_v \mathbf{1}_k \right) \\ &= \frac{1}{k}\left(\mathbf{1}_k^T f_v - \mathbf{1}_v^T X_v^T w_v \right) \end{aligned} \tag{6}$$

By setting the b derivative to 0, we have:

$$\begin{aligned} &\tilde{w}^T X 1_n + n b - f^T 1_n = 0 \\ &\Rightarrow b = \frac{1}{n}\left(f^T 1_n - \tilde{w}^T X 1_n \right) \\ &= \frac{1}{n}\left(1_n^T f - 1_n^T X^T \tilde{w} \right) \end{aligned} \tag{7}$$

Substituting Eq. (6) and Eq. (7) into Eq. (5), then setting its derivative on w_v to 0, we get the following formula:

$$w_v = Q_v^{-1}\left(\theta_v^r X_v H_k f_v + X_v X_v^T \tilde{w} \right) \tag{8}$$

where $H_k = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$, $Q_v = \theta_v^r X_v H_k X_v^T + X_v X_v^T + \beta V + \alpha I_d$ and $V = \left(W\left(W\right)^T\right)^{-1/2}$. By setting the derivative on $\tilde{w}$ to 0, we get:

$$\tilde{w} = A_v^{-1}\left(XH_n f + \theta_v^r X_v X_v^T Q_v^{-1} X_v H_k f_v\right) \tag{9}$$

where $A_v = XH_n X^T - X_v X_v^T Q_v^{-1} X_v X_v^T + X_v X_v^T + \beta I_d + \lambda_v X_v L_v X_v^T$ and $H_n = I_n - \frac{1}{n}1_n 1_n^T$. By setting its derivative for f_v to 0, we get:

$$f_v = B_v^{-1}\left(\theta_v^r H_k X_v^T Q_v^{-1} X_v X_v^T A_v^{-1} XH_n f\right) \tag{10}$$

where $B_v = \theta_v^r H_k - \theta_v^{2r} H_k X_v^T Q_v^{-1} X_v X_v^T A^{-1} X_v X_v^T Q_v^{-1} X_v \cdot H_k - \theta_v^{2r} H_k X_v^T Q_v^{-1} X_v H_k$.

By setting its derivative for f to 0, we get:

$$f = \begin{pmatrix} I - \theta_v^r H_n X^T A_v^{-1} X_v X_v^T Q_v^{-1} X_v H_k B_v^{-1} \theta_v^r H_k X_v^T Q_v^{-1} \\ X_v X_v^T A_v^{-1} XH_n - H_n X^T A_v^{-1} XH_n + H_n \end{pmatrix}^{-1} y \tag{11}$$

Optimize ${\theta_v}^r$ by fixing $b_v, w_v, f_v, \lambda_v, f, b$ and $\tilde{w}$.

After fixing $b_v, w_v, f_v, \lambda_v, f, b$ and $\tilde{w}$, the objective function in eq. (5) can be reformulated as

$$\begin{aligned} &\min_{\theta_v} \sum_{v=1}^{m}\left\{\theta_v^r \left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right\} \\ &s.t. \sum_{v=1}^{m}\theta_v = 1, \theta_v \in [0,1], \end{aligned} \tag{12}$$

By using the Lagrange multiplier δ, we convert the above problem into a Lagrange function as follows:

$$M\left(\theta_v, \delta\right) = \sum_{v=1}^{m}{\theta_v}^2\left(\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right) - \delta\left(\sum_{v=1}^{m}\theta_v - 1\right) \tag{13}$$

By setting its derivative for θ_i to 0, we get:

$$\begin{aligned} &\frac{\partial M}{\partial \theta_v} = 2\theta_v\left(\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right) - \delta \\ &\theta_v = \frac{\delta}{2}\left(\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right) \end{aligned} \tag{14}$$

Since $\sum_{v=1}^{m}\theta_v = 1$, we obtain:

$$\theta_v = \frac{\theta_v}{\sum_{v=1}^{m}\theta_v} = \frac{1/\left(\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right)}{\sum_{v=1}^{m} 1/\left(\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2\right)} \tag{15}$$

Optimize λ_v by fixing $b_v, w_v, f_v, \theta_v, f, b$ and $\tilde{w}$.

When fixing $b_v, w_v, f_v, \theta_v, f, b$ and $\tilde{w}$, the objective function in Eq. (5) is equivalent to:

$$\begin{aligned} &\min_{\lambda_v} tr\left(\tilde{w}^T\left(\sum_{v=1}^{m}\lambda_v X_v L_v X_v^T\right)\tilde{w}\right) + \mu\sum_{v=1}^{m}\lambda_v \log\lambda_v \\ &s.t. \sum_{v=1}^{m}\lambda_v = 1, \lambda_v \in [0,1] \end{aligned} \tag{16}$$

By using the Lagrange multiplier φ, we convert the above problem into a Lagrange function as follows:

$$\begin{aligned} L\left(\lambda_v, \varphi\right) = &tr\left(\tilde{w}^T\left(\sum_{v=1}^{m}\lambda_v X_v L_v X_v^T\right)\tilde{w}\right) \\ &+ \mu\sum_{v=1}^{m}\lambda_v \log\lambda_v - \varphi\left(\sum_{v=1}^{m}\lambda_v - 1\right) \end{aligned} \tag{17}$$

By setting its derivative for λ_v to 0, we have:

$$tr\left(\tilde{w}^T X_v L_v X_v^T \tilde{w}\right) + \mu\log\lambda_v + \mu - \varphi = 0$$

We thus obtain:

$$\lambda_v = \frac{\lambda_v}{\sum_{v=1}^{m}\lambda_v} = \frac{\exp\left(\left(-tr\left(\tilde{w}^T X_v L_v X_v^T \tilde{w}\right) - \mu\right)/\mu\right)}{\sum_{v=1}^{M}\exp\left(\left(-tr\left(\tilde{w}^T X_v L_v X_v^T \tilde{w}\right) - \mu\right)/\mu\right)} \tag{18}$$

Overall algorithm and convergence analysis

According to the above objective function optimization process, we summarize the following algorithm for LDG.

Below, we will demonstrate that the alternating optimization procedure converges to the optimal solution of $\{w_v\}_{v=1}^m$ corresponding to the optimization problem (5) according to Lemma 1.

> **Lemma 1.** For any invertible matrices M and $\tilde{V}$, the following inequality holds (Nie et al., 2010):

$$\frac{1}{2}tr\left(M\tilde{V}^{-\frac{1}{2}}\right) - tr\left(M^{\frac{1}{2}}\right) \geq \frac{1}{2}tr\left(\tilde{V}\tilde{V}^{-\frac{1}{2}}\right) - tr\left(\tilde{V}^{\frac{1}{2}}\right) \tag{19}$$

Next, we verify that the proposed iterative approach in Algorithm 1 can converge to the optimal solutions by the following theorem:

> **Theorem 1.** Algorithm 1 will monotonically decrease the objective of the problem in Eq. (5) in each iteration and will converge to the optimum of the problem.

Algorithm 1: Local domain generalization and adaptation

Input: Domain training dataset $\{(X_v, y_v), v = 1,\ldots,m\}$; the number of nearest neighbors k_1, k_2, and parameters α, β, and μ.

Initialization: Set $t = 0$, and initialize w_v^0, f_v^0, b_v^0, $\tilde{W}^0$, θ_v^0, λ_v^0, b^0, f^0 randomly, and set V_v^0 as identity matrix.

1: Construct the k-nearest neighbor graph and calculate $\{L_v\}_{v=1}^M$;

2:Compute H_k according to $H_k = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$;

3:Compute H_n according to $H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$;

4: For each v in $\{1,\ldots,m\}$

{

4.1: Let $t = 0$;

4.2: repeat

{

4.2.1:Compute λ_v^t according to Eq. (18);

4.2.2:Obtain θ_v^t by Eq. (15);

4.2.3: Compute Q_v^t as $Q_v^t = \left(\theta_v^t\right)^r X_v H_k X_v^T + X_v X_v^T + \beta V^t + \alpha I_d$;

4.2.4: Update A_v^t as $A_v^t = XH_nX^T - X_vX_v^T\left(A_v^t\right)^{-1}X_vX_v^T + X_vX_v^T + \beta I_d + \lambda_v^t X_v L_v X_v^T$;

4.2.5: Update B_v^t as $B_v^t = \left(\theta_v^t\right)^r H_k - \left(\theta_v^t\right)^{2r} H_k X_v^T\left(Q_v^t\right)^{-1}X_vX_v^T\left(A_v^t\right)^{-1}X_vX_v^T\left(Q_v^t\right)^{-1}X_vH_k - \left(\theta_v^t\right)^{2r} H_k X_v^T\left(Q_v^t\right)^{-1}X_vH_k$;

4.2.6: Compute f^0 according to Eq. (11);

4.2.7: Compute f_v^t according to Eq. (10);

4.2.8: Update $\ddot{U}_w^t$ as $\ddot{U}_w^t = A_v^{-1}\left(XH_nf^t + \left(\theta_v^t\right)^r X_vX_v^T\left(Q_v^t\right)^{-1}X_vH_kf_v^t\right)$;

4.2.9: Update w_v^t as $w_v^t = \left(Q_v^t\right)^{-1}\left(\left(\theta_v^t\right)^r X_vH_kf_v^t + X_vX_v^T\ddot{U}_w^t\right)$;

4.2.10: Update b_v^t as $b_v^t = \frac{1}{k}\left(\mathbf{1}_k^T f_v^t - \mathbf{1}_v^T X_v^T w_v^t\right)$;

4.2.11: Update b_v^t as $b^t = \frac{1}{n}\left(\mathbf{1}_n^T f^t - \mathbf{1}_n^T X^T\ddot{U}_w^t\right)$;

4.2.12: Update V_v^t;

4.2.13: Set $t = t + 1$;

} until $\left|\max \mathbb{C}_t - \min \mathbb{C}_t\right| / \max \mathbb{C}_t < 10^{-4}$;

4.3 Next v;

}

Output: Converged λ_v, θ_v, w_v, b_v, b_v, w, f, f_v.

Proof. For ease of representation, we denote the updated $b_v, w_v, f_v, \theta_v, \lambda_v$, b, and $\tilde{w}$ in each iteration as $b_v^l, w_v^l, f_v^l, \theta_v^l, \lambda_v^l, f^l, b^l$ and $\tilde{w}^l$, respectively. The inner loop to update in Step 2 of Algorithm 1 corresponds to the optimization of the following problem.

$$\begin{aligned}
&\min_{\substack{\theta_v, w_v, f, f_v \\ b_v, b, \tilde{w}, \lambda_v}} \sum_{v=1}^{n_+}\left\{\theta_v^r\left\|X_v^T w_v + b_v 1_k - f_v\right\|_2^2 + \alpha\left\|w_v\right\|_2^2\right\} \\
&+\sum_{v=1}^{n_+}\left\|X_v^T w_v - X_v^T\tilde{w}\right\|_2^2 + tr\left(\tilde{w}^T\left(\sum_{v=1}^{n_+}\lambda_v X_v L_v X_v^T\right)\tilde{w}\right) \\
&+\left\|X^T\tilde{w} + b1_n - f\right\|_2^2 + \left\|f - y\right\|_2^2 + \beta\left(\|W\|_* + \|\tilde{w}\|_2^2\right) \\
&+\mu\sum_{v=1}^{n_+}\lambda_v \log\lambda_v
\end{aligned} \tag{20}$$

According to the definitions of the matrix V, we obtain:

$$\begin{aligned}
&\min_{\substack{\theta_v, w_v, f, f_v \\ b_v, b, \tilde{w}, \lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{l+1}\right)^r\left\|X_v^T w_v^{l+1} + b_v^{l+1}1_k - f_v^{l+1}\right\|_2^2 + \alpha\left\|w_v^{l+1}\right\|_2^2\right\} \\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l+1} - X_v^T\tilde{w}^{l+1}\right\|_2^2 + tr\left(\left(\tilde{w}^{l+1}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l+1}X_v L_v X_v^T\right)\tilde{w}^{l+1}\right) \\
&+\left\|X^T\tilde{w}^{l+1} + b^{l+1}1_n - f^{l+1}\right\|_2^2 + \left\|f^{l+1} - y\right\|_2^2 + \beta\left(\left\|W^{l+1}\right\|_* + \left\|\tilde{w}^{l+1}\right\|_2^2\right) \\
&+\mu\sum_{v=1}^{M}\lambda_v^{l+1}\log\lambda_v^{l+1} \\
&\le \min_{\substack{\theta_v, w_v, f, f_v \\ b_v, b, \tilde{w}, \lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{l}\right)^r\left\|X_v^T w_v^{l} + b_v^{l}1_k - f_v^{l}\right\|_2^2 + \alpha\left\|w_v^{l}\right\|_2^2\right\} \\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l} - X_v^T\tilde{w}^{l}\right\|_2^2 + tr\left(\left(\tilde{w}^{l}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l}X_v L_v X_v^T\right)\tilde{w}^{l}\right) \\
&+\left\|X^T\tilde{w}^{l} + b^{l}1_n - f^{l}\right\|_2^2 + \left\|f^{l} - y\right\|_2^2 + \beta\left(\left\|W^{l}\right\|_* + \left\|\tilde{w}^{l}\right\|_2^2\right) \\
&+\mu\sum_{v=1}^{M}\lambda_v^{l}\log\lambda_v^{l}
\end{aligned} \tag{21}$$

Eq. (21) is equivalent to the following form:

$$
\begin{aligned}
&\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l+1}\right)^r \left\|X_v^T w_v^{l+1}+b_v^{l+1}\mathbf{1}_k-f_v^{l+1}\right\|_2^2+\alpha\left\|w_v^{l+1}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l+1}-X_v^T\tilde{w}^{l+1}\right\|_2^2+tr\left(\left(\tilde{w}^{l+1}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l+1}X_vL_vX_v^T\right)\tilde{w}^{l+1}\right)\\
&+\left\|X^T\tilde{w}^{l+1}+b^{l+1}1_n-f^{l+1}\right\|_2^2+\left\|f^{l+1}-y\right\|_2^2+\beta\left\|\tilde{w}^{l+1}\right\|_2^2\\
&+\mu\sum_{v=1}^{M}\lambda_v^{\,l+1}\log\lambda_v^{\,l+1}+\beta tr\left(W^{l+1}\left(W^{l+1}\right)^T V^{l+1}\right)-\frac{\beta}{2}\mathrm{tr}\left(\left(W^{l+1}\left(W^{l+1}\right)^T\right)^{\frac{1}{2}}\right)\\
&+\frac{\beta}{2}\mathrm{tr}\left(\left(W^{l+1}\left(W^{l+1}\right)^T\right)^{\frac{1}{2}}\right)\\
&\le\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l}\right)^r \left\|X_v^T w_v^{l}+b_v^{l}\mathbf{1}_k-f_v^{l}\right\|_2^2+\alpha\left\|w_v^{l}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l}-X_v^T\tilde{w}^{l}\right\|_2^2+tr\left(\left(\tilde{w}^{l}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l}X_vL_vX_v^T\right)\tilde{w}^{l}\right)\\
&+\left\|X^T\tilde{w}^{l}+b^{l}1_n-f^{l}\right\|_2^2+\left\|f^{l}-y\right\|_2^2+\beta\left\|\tilde{w}^{l}\right\|_2^2\\
&+\mu\sum_{v=1}^{M}\lambda_v^{\,l}\log\lambda_v^{\,l}+\beta tr\left(W^{l}\left(W^{l}\right)^T V^{l}\right)-\frac{\beta}{2}\mathrm{tr}\left(\left(W^{l}\left(W^{l}\right)^T\right)^{\frac{1}{2}}\right)\\
&+\frac{\beta}{2}\mathrm{tr}\left(\left(W^{l}\left(W^{l}\right)^T\right)^{\frac{1}{2}}\right)
\end{aligned}
\tag{22}
$$

Since $V^l=\frac{1}{2}\left(W^l\left(W^l\right)^T\right)^{-\frac{1}{2}}$ and according to Lemma 1, we obtain:

$$
\begin{aligned}
&\beta tr\left(W^{l+1}\left(W^{l+1}\right)^T V^{l+1}\right)-\beta tr\left(\left(W^{l+1}\left(W^{l+1}\right)^T\right)^{\frac{1}{2}}\right)\\
&\ge\beta tr\left(W^{l}\left(W^{l}\right)^T V^{l}\right)-\beta tr\left(\left(W^{l}\left(W^{l}\right)^T\right)^{\frac{1}{2}}\right)
\end{aligned}
\tag{23}
$$

Subtracting (23) from (22), we have:

$$
\begin{aligned}
&\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l+1}\right)^r \left\|X_v^T w_v^{l+1}+b_v^{l+1}1_k-f_v^{l+1}\right\|_2^2+\alpha\left\|w_v^{l+1}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l+1}-X_v^T\tilde{w}^{l+1}\right\|_2^2\\
&+tr\left(\left(\tilde{w}^{l+1}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l+1}X_vL_vX_v^T\right)\tilde{w}^{l+1}\right)\\
&+\left\|X^T\tilde{w}^{l+1}+b^{l+1}1_n-f^{l+1}\right\|_2^2+\left\|f^{l+1}-y\right\|_2^2+\beta\left\|\tilde{w}^{l+1}\right\|_2^2\\
&+\mu\sum_{v=1}^{M}\lambda_v^{\,l+1}\log\lambda_v^{\,l+1}+\frac{\beta}{2}tr\left(\left(W^{l+1}\left(W^{l+1}\right)^T\right)^{\frac{1}{2}}\right)\\
&\le\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l}\right)^r \left\|X_v^T w_v^{l}+b_v^{l}1_k-f_v^{l}\right\|_2^2+\alpha\left\|w_v^{l}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l}-X_v^T\tilde{w}^{l}\right\|_2^2+tr\left(\left(\tilde{w}^{l}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l}X_vL_vX_v^T\right)\tilde{w}^{l}\right)\\
&+\left\|X^T\tilde{w}^{l}+b^{l}1_n-f^{l}\right\|_2^2+\left\|f^{l}-y\right\|_2^2+\beta\left\|\tilde{w}^{l}\right\|_2^2+\mu\sum_{v=1}^{M}\lambda_v^{\,l}\log\lambda_v^{\,l}+\frac{\beta}{2}tr\left(\left(W^{l}\left(W^{l}\right)^T\right)^{\frac{1}{2}}\right)
\end{aligned}
\tag{24}
$$

The above formula is equivalent to:

$$
\begin{aligned}
&\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l+1}\right)^r \left\|X_v^T w_v^{l+1}+b_v^{l+1}1_k-f_v^{l+1}\right\|_2^2+\alpha\left\|w_v^{l+1}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l+1}-X_v^T\tilde{w}^{l+1}\right\|_2^2\\
&+tr\left(\left(\tilde{w}^{l+1}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l+1}X_vL_vX_v^T\right)\tilde{w}^{l+1}\right)+\left\|X^T\tilde{w}^{l+1}+b^{l+1}1_n-f^{l+1}\right\|_2^2\\
&+\left\|f^{l+1}-y\right\|_2^2\\
&+\beta\left\|\tilde{w}^{l+1}\right\|_2^2+\mu\sum_{v=1}^{M}\lambda_v^{\,l+1}\log\lambda_v^{\,l+1}+\beta\left\|W^{l+1}\right\|_*\\
&\le\min_{\substack{\theta_v,w_v,f_v,f_v\\ b_v,b,\tilde{w},\lambda_v}} \sum_{v=1}^{M}\left\{\left(\theta_v^{\,l}\right)^r \left\|X_v^T w_v^{l}+b_v^{l}1_k-f_v^{l}\right\|_2^2+\alpha\left\|w_v^{l}\right\|_2^2\right\}\\
&+\sum_{v=1}^{M}\left\|X_v^T w_v^{l}-X_v^T\tilde{w}^{l}\right\|_2^2+tr\left(\left(\tilde{w}^{l}\right)^T\left(\sum_{v=1}^{M}\lambda_v^{l}X_vL_vX_v^T\right)\tilde{w}^{l}\right)\\
&+\left\|X^T\tilde{w}^{l}+b^{l}1_n-f^{l}\right\|_2^2+\left\|f^{l}-y\right\|_2^2+\beta\left\|\tilde{w}^{l}\right\|_2^2+\mu\sum_{v=1}^{M}\lambda_v^{\,l}\log\lambda_v^{\,l}+\beta\left\|W^{l}\right\|_*
\end{aligned}
\tag{25}
$$

Therefore, we have proved the theorem. Because of the updating rule in Algorithm 1, the objective function shown in (5) monotonically decreases, and it is easy to see that the algorithm converges.

Target inference

After training the LDG, we get *m* local classifiers. In the following sections, we will separately propose ways to effectively use these learned classifiers in two cases.

1. LDG: The first is a domain generalization scenario where the target domain samples are not available during training. The other is the domain adaptation scenario with a specific target domain in which we have unlabeled data in it during the training process. In the domain generalization scenario, under the premise that we have no prior information about the target domain, we can only fuse the *m* local classifiers to achieve the prediction of the test sample by assigning different weights. Given a target sample *x*, the predictive label *y* can be obtained by the following formula.

$$
y=\sum_{v=1}^{m}\theta_v^2 f_v\left(w_v,X_v\right)=\sum_{v=1}^{m}\theta_v^2\left(x^T w_v+b_v\right)
\tag{26}
$$

2. LDA: When there is unlabeled data in the target domain, we can assign different weights to each local classifier by measuring the similarity between the target domain and each locality in the source domain to achieve a better prediction effect. In other words, when a certain local domain is closer to the target domain, we should assign a higher weight to the classifier trained on this subdomain, and vice versa.

Given a set of target domain samples $X=\left\{x_1,x_2,\ldots,x_K\right\}$, where K is the number of samples in the target domain. By measuring the distance between the training sample and the target domain by the Maximum Mean Discrepancy (MMD), we get the following formula:

$$
\Psi^v=Dist\left(X_v,X\right)=\frac{1}{k}\sum_{i=1}^{k}\phi\left(x_i^v\right)-\frac{1}{K}\sum_{j=1}^{K}\phi\left(x_j\right)_{H_k}
\tag{27}
$$

where X_v, X are the v-th local source domain and target domain datasets respectively, and $Dist(X_v, X)$ represents the distribution distance of X_v and X, and H_K denotes a regenerative kernel Hilbert space. $\phi(\cdot)$ is a Gaussian kernel nonlinear feature mapping function. Using MMD we can get the weight of each local classifier by:

$$\zeta_v = \frac{\exp\left(-\Psi^v\right)}{\sum_{v=1}^{m}\exp\left(-\Psi^v\right)}, v = 1, 2, \cdots, m \tag{28}$$

Then we can predict the test sample x_j by the following formula:

$$y_j = \sum_v \zeta_v \left(x_j^T w_v + b_v\right) \tag{29}$$

Experimental results

In this section, we will conduct comprehensive experiments to validate the effectiveness of our method compared with several state-of-the-art ones.

Benchmark datasets

Two widely used benchmark databases, i.e., SEED (Zheng and Lu, 2015) and DEAP (Koelstra et al., 2012), are adopted for systematic experiments of EEG-based emotion recognition (Dan et al., 2021; Tao et al., 2022). More detailed descriptions of these two benchmarks can be found in Lan et al. (2018). As reported by references (Zhong P. et al., 2020; Zhong Q. et al., 2020) and (Lan et al., 2018), some obvious differences between these two benchmarks are that they may be sampled from multiple different sources such as different sessions, subjects, experimental schemes, EEG devices, and emotional stimuli, etc. Following the same practice in literature (Shi et al., 2013; Zheng et al., 2015; Chai et al., 2016, 2017; Zheng and Lu, 2016; Lan et al., 2018; Zhong P. et al., 2020; Zhong Q. et al., 2020; Tao and Dan, 2021; Tao et al., 2022) for domain adaptation emotion recognition, differential entropy (DE; Lan et al., 2018; Zhong P. et al., 2020; Zhong Q. et al., 2020) is adopted as the data feature in our experimental settings.

Baselines and protocol

Baselines

As a DG method, we compare our method with several representative domain generalization/adaptation methods, which can be summarized into the following two groups (here we only report the better models):

1. Shallow learning methods: Undo-Bias (Khosla et al., 2012), UML (Fang et al., 2013), DICA (Muandet et al., 2013), LRE-SVM (Xu et al., 2014), and SCA (Ghifary et al., 2017);
2. Deep learning methods: Deep subdomain adaptation network (DSAN; Zhu et al., 2020), Deep domain generalization with structured low-rank constraint (DDG) (Ding et al., 2018a,b,c), deep domain confusion (DDC) (Tzeng et al., 2014), domain adversarial neural networks (DANNs) (Ganin et al., 2016), contrastive adaptation network (CAN) (Kang et al., 2022), and deep CORAL (Sun and Saenko, 2016).

Training protocol

For all datasets, we only exploit the source samples for training. We use support vector machine (SVM) as the base classifier for DICA and SCA. The tunable hyper-parameters are selected according to labels from the source domain. We adopt the Gaussian kernel with a kernel bandwidth σ computed by the median heuristic as the kernel function for the kernel-based methods. For a fair comparison, all deep learning baselines use the same architecture (ResNet101; He et al., 2016). That is, for deep domain generalization on the EEG dataset, we employed the Resnet101 architecture to extract the training features. We fine-tune all convolutional and pooling layers from pre-trained models and train the classifier layer via back-propagation. For multi-class classification of emotion recognition, we employ the "One vs. Rest" strategy to train our method (Zhang et al., 2020b).

Parameter setting

There are several vital parameters such as μ, α, and β that need to be determined beforehand in our objective (5) since they are employed to balance the importance of structure characterization and regularizers. Considering that parameter determination is a yet unaddressed open issue in the field of machine learning, we determine them by grid search in a heuristic way (Nie et al., 2010; Long et al., 2014b; Tao et al., 2022). Concretely, these regularization parameters are tuned from $\{10^{-4}, 10^{-3}, \ldots, 10^{3}, 10^{4}\}$. Since no target labels are available for DG, it is impossible to conduct a standard cross-validation. Hence, we perform p-fold cross-validation on the labeled source subdomains, namely, calculating the averaged accuracy on each subdomain fold while exploiting the other $p-1$ subdomain folds for training. Moreover, for constructing the nearest neighbor graph in LDG, we search the optimal neighbor number k (including k_1 and k_1) in the grid $\{3, 5, 7, 9, 11, 13\}$, and then report the top-one recognition accuracy from the best parameter configuration. For the kernel learning scenarios, the Gaussian kernel, i.e., $K_{i,j} = \exp\left(-\sigma\left\|x_i - x_j\right\|^2\right)$, is used as the default kernel function, where σ is set to $1/d$ (d is the feature dimension).

Inter-subject domain generalization

Note that different subjects even from the same dataset still have different EEG feature distributions due to their characteristics. We therefore conduct the so-called leave-one-out cross-validation strategy conducted also in Lan et al. (2018) and Tao et al. (2022) to evaluate the emotion recognition performance. That is, one subject remains to be the target domain, and others from the dataset are constructed as the source domain. In this scenario, we follow the same setting as (Lan et al., 2018; Tao and Dan, 2021; Tao et al., 2022) to evaluate our method compared with other state-of-the-art methods on SEED and DEAP, respectively.

Each subject from DEAP includes 180 samples belonging to three categories, i.e., 60 samples per class. Each subject from SEED

TABLE 2 Inter-subject recognition accuracy (mean % and STD %).

Method		DEAP		SEED							
				Session I		Session II		Session III		Average	
		Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Shallow methods	Undo-Bias	60.36	3.41	69.41	5.44	65.79	2.24	72.64	5.10	69.28	4.26
	UML	62.18	4.09	72.57	6.27	67.58	1.75	71.17	3.68	70.44	3.90
	DICA	65.33	6.22	73.12	6.86	65.06	6.28	73.38	7.19	70.52	6.78
	LRE-SVM	68.20	2.12	77.50	3.29	70.11	5.44	77.45	4.53	75.02	4.42
	SCA	66.05	4.26	75.23	5.17	69.14	6.20	74.23	6.07	72.87	5.81
	LDG	71.51	3.14	78.92	5.65	70.88	5.72	78.93	5.38	76.24	5.58
Deep methods	DDG	77.68	3.33	84.92	6.42	74.29	7.45	82.33	8.11	80.51	7.33
	DDC	74.87	6.28	79.43	7.13	72.16	6.11	80.07	7.66	77.22	6.97
	DANN	75.34	7.11	82.51	6.49	73.77	7.59	83.62	6.51	79.97	6.86
	DSAN	78.44	4.15	84.50	6.18	74.58	6.33	84.10	6.12	81.06	6.21
	CORAL	74.08	3.58	80.42	4.20	71.54	5.49	81.00	5.00	77.65	4.90
	CAN	78.43	6.10	85.77	7.31	74.12	7.50	85.39	7.40	81.76	7.40
	LDG + Resnet101	77.62	5.37	85.42	5.72	74.68	5.19	86.05	6.82	82.05	5.91
Upp Bnd of Chn Lvl (UBCL)		38.85		34.58		34.65		34.60		34.61	

Bold denotes the best recognition rates.

contributes 2,775 samples, i.e., 925 samples per class and per session. Following the same strategy adopted by Chai et al. (2016), Zheng and Lu (2016), and Chai et al. (2017), we randomly sampled 1/10 of the training data (3,885 samples contributed by 14 subjects) from SEED in each experiment due to the large number of training samples. To cover the whole training dataset, we randomly extracted 10 training sets from SEED and thus conducted each experimental procedure 10 times. The final result was averaged over these 10 runs. We compared the performance of our LDG with several state-of-the-art DG approaches. The mean recognition accuracies of LDG compared with the baselines on two benchmark datasets are recorded in Table 2.

As is known, when the size of training data increases to infinity, the theoretical performance (about 33.33%) of the random prediction can be approximately approached by the real chance level (Lan et al., 2018). When there are finite samples, we obtain the empirical chance level by repeating the trials with the samples in question equipped with randomized class labels (Lan et al., 2018). For comparison, we also provided the upper bound of chance levels (UBCL) with a 95% confidence interval in this table.

Comparison with shallow methods

As observed from Table 2, the mean performance of all methods on two datasets has significantly exceeded UBCL at a 5% significance level. This indicates the imperative importance of inter-subject domain generalization due to the intrinsic existence of distribution divergence among different subjects. Compared with other shallow learning methods, our method LDG undoubtedly obtains the best mean performance (75.06% ± 4.97) in all cases, which is followed by LRE-SVM (73.32% ± 3.85). This may be attributed to the subdomain learning technologies in LDG and LRE-SVM. Our method LDG unsurprisingly achieved more performance gains than LRE-SVM on both DEAP and SEED. The multi-source generalization method SCA and DICA are found to be more effective than Undo-bias and UML. The experimental results in Table 2 show that while the relative improvement over vanilla DA/DG methods is significant (t-test, value of $p > 0.05$), the absolute accuracy is still rather low, which suggests that there still exists adverse impact incurred by substantial distribution discrepancies between different subjects.

An interesting result that can be observed from Table 2 is that all methods demonstrate better performance on SEED than on DEAP. The same observation has also been reported in Lan et al. (2018) and Tao and Dan (2021). A possible explanation for this result might be that there exist large discrepancies among different subjects, and the samples are distributed more "orderly" in their original feature space on SEED than that on DEAP (Mansour et al., 2009), thus leading to better alignment on SEED in some kernel space. That is, larger discrepancies among different subjects from DEAP could degrade the recognition accuracy to some extent (Mansour et al., 2009; Lan et al., 2018).

Comparison with deep methods

Following the same settings in Donahue et al. (2014) and Zhou et al. (2022), our method LDG also can be easily trained with the deeply extracted features via the classical deep models such as VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016). Specifically, one can fine-tune some pre-trained deep models (e.g., Resnet101; He et al., 2016) through the source domain, and then extract the deep features from EEG signals. Finally, the recognition model can be learned using these deep representations.

In this part, we will particularly evaluate our method LDG with deeply extracted features by comparing it with several recently proposed deep adaptation models. We additionally denote our

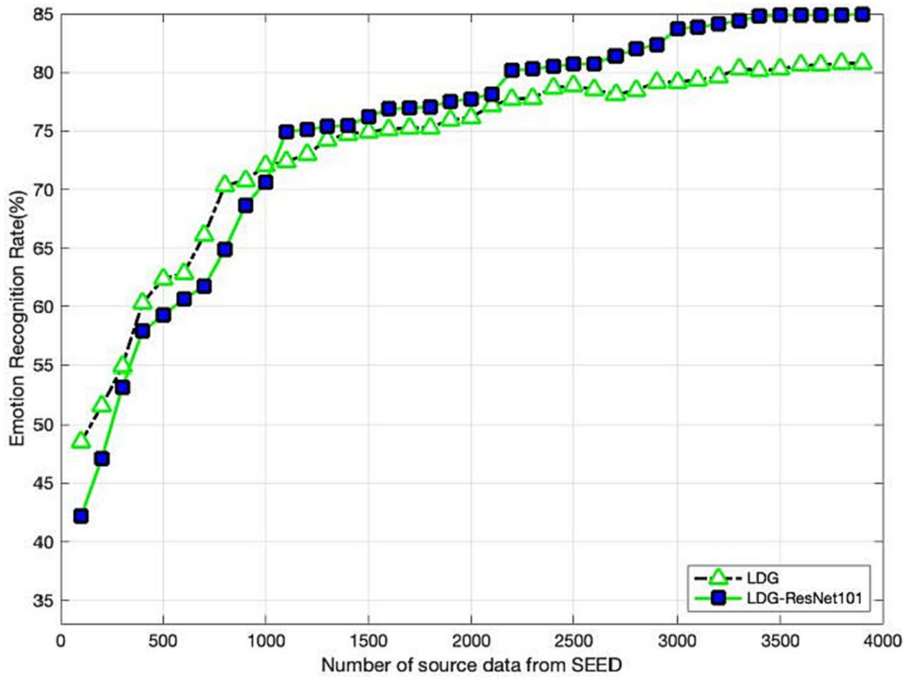


FIGURE 2
Recognition accuracy with varying sizes of source samples on SEED.

method with deep features as LDG + ResNet101. As for other deep benchmarks, their opened source codes are directly borrowed to fine-tune the pre-trained models adopted in their works, respectively. Different from these deep adaptation models, which typically pursue gaining certain domain-invariant representations, our proposed method explores optimizing a domain-invariant recognition model with strong generalization ability from the single source domain to the unseen target. We expect our method leveraging the deeply extracted features can further upgrade the recognition performance with the proposed subdomain generalization strategy.

As shown in Table 2, all of the deep methods obviously outperform the shallow ones. This indicates the advantage of deep features due to their more discriminative representations. As expected, LDG + ResNet101 also obtains the best or comparable recognition performance compared with other deep adaptation methods, followed by CAN and DSAN. This may be partly attributed to the classification-level modeling in our LDG, where most of the local discriminative structures have been preserved by the guidance of subdomain construction. In some scenarios, shown in Table 2, LDG + ResNet101 even achieves the top-one accuracy, which verifies that the proposed LDG can become an effective surrogate to the deep adaptation model by exploiting the deeply extracted features from some pre-trained deep models.

Sample size impact

Figure 2 clearly reports the impact on the performance with different sizes of source on SEED, where the source size varies from 100 to 3,800. We can observe that our methods LDG and LDG + ResNet101 manifest the same trends of upgrade in the curves. As expected, larger source samples are beneficial to improve the recognition performance of our methods. It is worth noting that the performance of LDG can be smoothly and steadily improved with the increase of the source size, while LDG + ResNet101 can achieve significant performance when the source samples are relatively large, e.g., larger than 1,100. When the number of source samples increases to 3,500, LDG and LDG-ReNet101 asymptotically approach their equilibrium states.

Multiple kernel selection

As an open problem, how to choose an effective kernel is a challenge for learning a kernel machine. Fortunately, the previously proposed multiple kernel learning (MKL) trick can be adapted to overcome this confusion. In the sequence, we further evaluate the performance improvement in our method via introducing MKL (denoted by LDG-mkl for short) for each subdomain, in which a new feature space spanned by multiple kernel projections will be first constructed. Specifically, given an empirical kernel function set $\{\phi_a\}_{a=1}^{\mho}$, we, respectively, project them into $\mho$ different spaces, and then construct an orthogonally integrated feature space by horizontally concatenating these spaces. In this concatenated space, the projected features can be denoted by $\tilde{\phi}(x_i) = \left[\phi_1(x_i)^T, \phi_2(x_i)^T, \ldots, \phi_{\mho}(x_i)^T\right]^T \in \mathbb{R}^{\mho n_a}$, where $x_i \in X_a$, and then the kernel matrix can be easily deduced as $K_{new} = \left[\tilde{K}_1; \tilde{K}_2; \ldots; \tilde{K}_{\mho}\right]$, where $\tilde{K}_i$ is the i-th kernel matrix from the $\mho$ feature spaces. Following the same strategy in Long et al. (2015), besides the above-used Gaussian kernel, we additionally introduce another three kernel functions including inverse

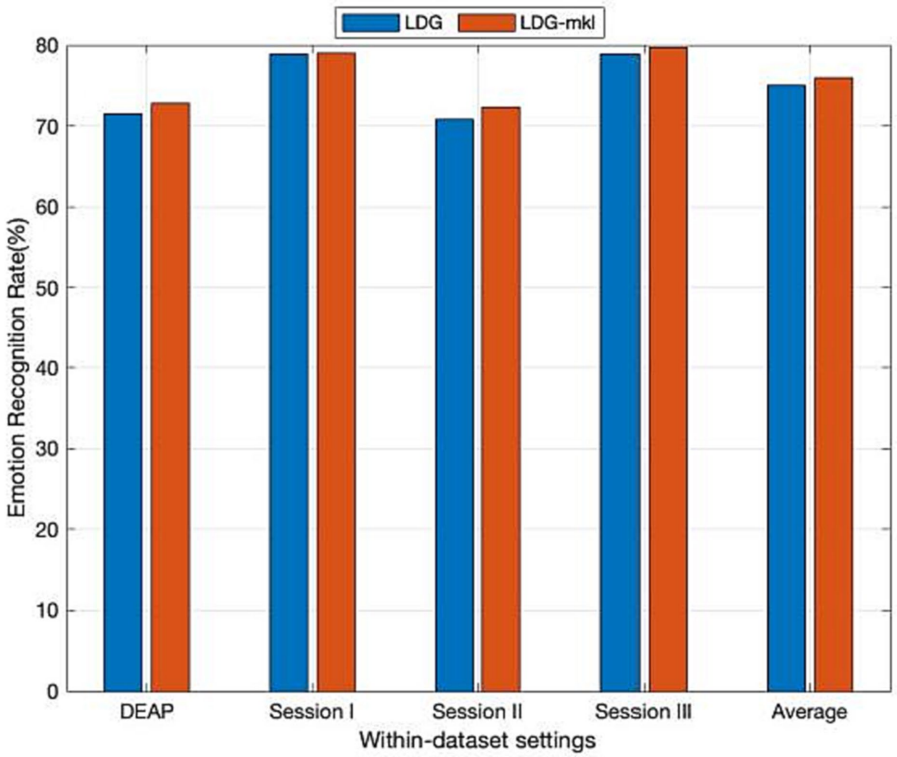


FIGURE 3
Domain adaptation emotion recognition on within-dataset with multi-kernel learning (SI: Session I, SII: Session II, SIII: Session III).

square distance kernel function, Laplacian kernel function, and inverse distance kernel function, which are, respectively, denoted as $K_{ij} = 1/\left(1+\sigma\left\|x_i - x_j\right\|^2\right)$, $K_{ij} = \exp\left(-\sqrt{\sigma}\left\|x_i - x_j\right\|\right)$, and $K_{ij} = 1/\left(1+\sqrt{\sigma}\left\|x_i - x_j\right\|\right)$. The observed mean experimental results from Figure 3 prove that LDG-mkl can boost the performance of LDG with a single kernel. This also verifies that the performance improvement in the kernel machines can be attributed to the diversities of multiple kernel functions.

Cross-dataset domain generalization

In this subsection, we further evaluate the broad and consistent generalization capacity of our LDG method on cross-dataset emotion recognition. Intuitively speaking, cross-data generalization must be more challenging than cross-subject generalization due to the significant difference between datasets.

Following the same settings in Tao and Dan (2021) and Tao et al. (2022), for robust cross-dataset generalization, the 32 shared channels by SEED and DEAP are employed to support a common feature space of 160 dimensions. In this case, several cross-dataset generalization settings can be made up, i.e., $DEAP \rightarrow SI$, $DEAP \rightarrow SII$, $DEAP \rightarrow SIII$, $SI \rightarrow DEAP$, $SEED\ II \rightarrow DEAP$, and $SIII \rightarrow DEAP$, where "$x \rightarrow y$" means domain generalization from the dataset x into the dataset y, and SI, SII, and SIII are, respectively, denoted as the Session I, Session II, and Session III from SEED. When DEAP is regarded as the source dataset, 2,520 data are sampled from DEAP and 2,775 data taken as the target datasets are, respectively, sampled from three different sessions (SI, SII, and SIII) of SEED. When each session of SEED is taken as the source dataset, we resample 41,625 data from it as a training set and 180 samples from DEAP regarded as the target dataset. We report the mean generalization results on six cross-dataset in Table 3.

It can be seen from the experimental results in Table 3 that the average performance of each method on the cross-dataset is slightly worse than that in the within-dataset. This confirms that the distribution difference between the two datasets is greater than that between the two subjects. The superiority of subdomain generalization will be reflected in this scenario because subdomains can potentially alleviate the distribution diversity in cross-datasets when the target dataset is unavailable in the phase of training. This can also be proved by the observation from Table 3, where our method LDG outperforms other shallow methods in almost all cases. Although SCA occasionally achieves the best performance in one setting (SI → DEAP), our LDG method still achieves the top-one performance in other cases. In deep learning scenarios, all methods still undoubtedly outperform their shallow counterparts, which can be attributed to the advantage of deep feature representations. It is worth noting that our deep method LDG + Resnet101 also obtains the best or comparable recognition performance compared with other deep adaptation models. This once again proves the importance of the classification-level constraint in LDG.

Regarding the previously reported results in Yang et al. (2007), Tommasi et al. (2014), Tao et al. (2017, 2019, 2022), Ding et al.

TABLE 3 Domain adaptation emotion recognition on cross-dataset.

Methods		DEAP → SI	DEAP → SII	DEAP → SIII	SI → DEAP	SII → DEAP	SIII → DEAP
Shallow methods	Undo-Bias	44.35	49.72	43.71	42.57	41.99	42.51
	UML	45.63	49.98	49.67	44.91	42.48	43.53
	DICA	47.35	52.68	52.11	43.34	44.90	42.46
	LRE-SVM	50.48	56.46	57.11	46.34	47.20	47.46
	SCA	48.89	54.35	54.65	**46.73**	45.36	44.67
	LDG	**52.62**	**57.66**	**57.83**	45.60	**47.89**	**49.76**
Deep methods	DDG	62.40	64.92	73.92	64.29	54.29	53.33
	DDC	60.89	62.43	69.43	62.16	52.16	50.07
	DANN	61.08	62.51	72.51	63.77	53.77	52.62
	DSAN	63.28	64.50	**74.50**	64.58	55.58	54.10
	CORAL	60.15	60.42	70.42	61.54	52.54	51.00
	CAN	64.22	65.77	75.77	66.12	**57.12**	55.39
	LDG+ Resnet101	**62.62**	**65.81**	74.42	**66.86**	55.68	**56.18**

Bold denotes the best recognition rates.

TABLE 4 Multi-dataset generalization (SI: Session I, SII: Session II, SIII: Session III).

Methods	{DEAP,SII,SIII} → SI	{DEAP,SI,SIII} → SII	{DEAP,SI,SII} → SIII	{SI,SII,SIII} → DEAP	{SI,SII} → DEAP	{SI,SIII} → DEAP
Undo-Bias	46.16	51.32	45.11	43.84	40.68	41.76
UML	44.06	50.10	51.21	46.01	42.90	43.85
DICA	49.28	52.94	54.06	46.62	45.39	44.63
LRE-SVM	47.17	57.30	59.30	**50.77**	46.50	49.06
SCA	52.33	**57.66**	57.29	48.68	**47.72**	48.93
LDG	**52.76**	**57.66**	**61.43**	49.34	47.03	**49.48**

Bold denotes the best recognition rates.

(2018a,b,c), and Tao and Dan (2021), multi-dataset adaptation can be improved by ensemble multiple auxiliary datasets. Please note that the scalability challenge could be incurred in case of multi-dataset generalization in that multi-dataset learning could bring the so-called "negative transfer" problem (Rosenstein et al., 2005), an open issue that exists in vanilla multi-source DA (Li J. et al., 2019; Chen et al., 2021; Tao and Dan, 2021). Therefore, we particularly evaluate the scalability of the proposed method by leveraging multiple source datasets for cross-dataset generalization. We report the average performance in Table 4 on all source datasets for the single-source methods including our LDG as well as LRE-SVM, Undo-Bias, and UML.

As shown in Table 4, due to the significant distribution differences among different source datasets, it is difficult for the single-source methods to generalize to unseen target domains in multi-source datasets. Therefore, the results in Table 4 indicate that these methods are generally inferior to other multi-source fusion methods. In some scenarios, they even exhibit a performance degradation trend as the number of source domains increases, indicating the "negative transfer" phenomenon (Rosenstein et al., 2005). Another interesting observation in Table 4 is that all multi-source methods achieve slight improvements by utilizing multiple sources as opposed to bridging only a single source (i.e., cross-dataset settings) as the number of source domains increases. This demonstrates the benefits of using multiple sources to enhance identification performance. In addition, SCA and DICA outperform other methods by conquering top-level performance as their designed weights are used to differentially screen the best sources. In some cases, our LDG method achieves more benefits than SCA. One possible explanation is that the discriminative information shared among sub-domain models in LDG is advantageous for multi-source generalization.

Convergence

Since our LDG is an iterative algorithm, it is crucial to evaluate its convergence. We evaluate the convergence of the LDG algorithm by conducting several experiments for emotion recognition in three settings such as cross-subject within DEAP, DEAP→SI, and {SI, SII, SIII} → DEAP. We plotted the mean experimental results in Figure 4. The curves in this figure show that the proposed algorithm has a certain asymptotic convergence. The objective values of LDG usually converge in less than 30 iterations. We also observed a similar phenomenon from other recognition tasks with different cross-subject/cross-dataset settings.

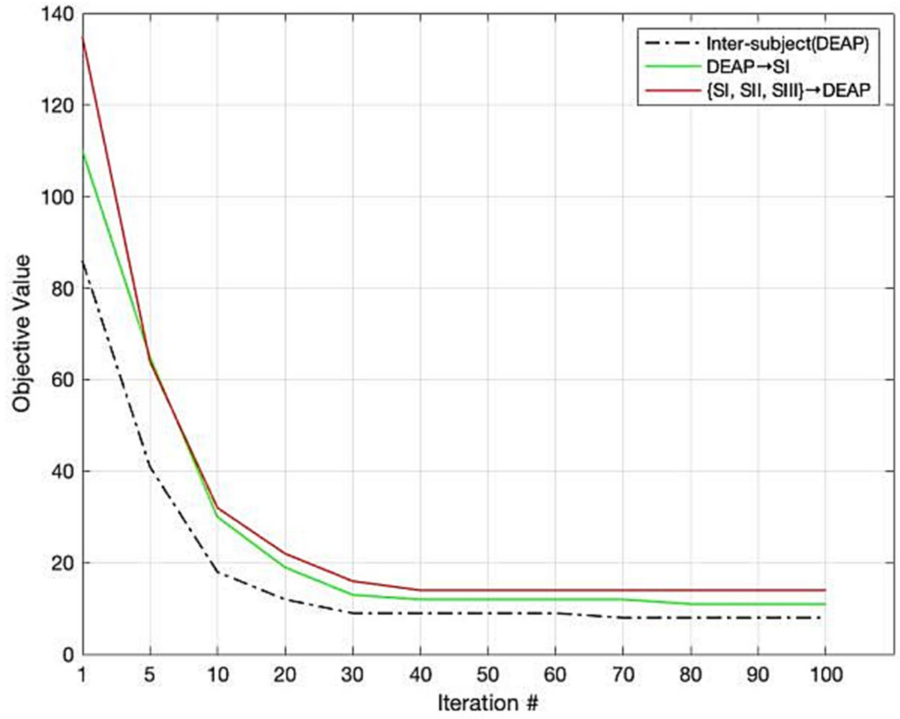


FIGURE 4
Convergence of LDG.

Conclusion

To deal with cross-subject/dataset EEG-based emotion recognition tasks, we proposed a local domain generalization (LDG) framework. In multiple subdomain spaces, LDG aims at transferring local knowledge into target learning mainly by leveraging correlation knowledge among subdomain models via low-rank constraint on the local models, which discriminatively screens unimportant prior evidence in subdomains. The comprehensive experiments performed on two public datasets verify the effectiveness of LDG in dealing with cross-subject/dataset emotion recognition. In most scenarios, our LDG and LDG + Resnet101 obtain the best results or comparable performance concerning several representative baselines.

Since the implementation of the LDG algorithm needs an iterative optimization procedure, how to improve the efficiency of LDG and seek a more efficient algorithm would be an issue worthy of further study in our future research. The unreliable and misleading pseudo-label strategy may be another potential problem in our LDG. Consequently, our successive work would be to explore seamlessly incorporating target labels into the framework of LDG.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: in our research, the datasets DEAP and SEED can be, respectively, accessed from http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html and http://bcmi.sjtu.edu.cn/~seed.

Author contributions

DZ extensively conducted all experiments in the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Ningbo Natural Science Foundation project (No. 2022J180).

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

References

Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Trans. PAMI* 32, 770–787. doi: 10.1109/TPAMI.2009.57

Chai, X., Wang, Q., Zhao, Y., Li, Y., Liu, D., Liu, X., et al. (2017). A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition. *Sensors* 17:1014. doi: 10.3390/s17051014

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., and Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Comput. Biol. Med.* 79, 205–214. doi: 10.1016/j.compbiomed.2016.10.019

Chang, H., Zong, Y., Zheng, W., Tang, C., Zhu, J., and Li, X. (2021). Depression assessment method: an EEG emotion recognition framework based on spatiotemporal neural network. *Front. Psychiatry* 12:837149. doi: 10.3389/fpsyt.2021.837149

Chang, H., Zong, Y., Zheng, W., Xiao, Y., Wang, X., Zhu, J., et al. (2023). EEG-based major depressive disorder recognition by selecting discriminative features via stochastic search. *J. Neural Eng.* 20:026021. doi: 10.1088/1741-2552/acbe20

Chen, H., Jin, M., Li, Z., Fan, C., Li, J., and He, H. (2021). MS-MDA: multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition. *Front. Neurosci.* 15:778488. doi: 10.3389/fnins.2021.778488

Chen, B., Lam, W., Tsang, I. W., and Wong, T. L. (2013). Discovering low-rank shared concept space for adapting text mining models. *IEEE Transac. Pattern Analy. Mach. Intellig.* 35, 1284–1297. doi: 10.1109/TPAMI.2012.243

Dan, Y., Tao, J., Fu, J., and Zhou, D. (2021). Possibilistic clustering-promoting semi-supervised learning for EEG-based emotion recognition. *Front. Neurosci. Brain Imag. Methods* 15:690044. doi: 10.3389/fnins.2021.690044

Ding, Z., Nasser, M. N., and Fu, Y. (2018b). Semi-supervised deep domain adaptation via coupled neural networks. *IEEE Trans. Image Process.* 27, 5214–5224. doi: 10.1109/TIP.2018.2851067

Ding, Z., Shao, M., and Fu, Y. (2018c). Incomplete multisource transfer learning. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 310–323. doi: 10.1109/TNNLS.2016.2618765

Ding, Z, Sheng, L, Ming, S, and Fu, Y. (2018a). Graph adaptive knowledge transfer for unsupervised domain adaptation. 15th European Conference (ECCV 2018), Munich, Germany, September 8–14, 2018, Springer, Cham.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). DeCAF: a deep convolutional activation feature for generic visual recognition. *PMLR* 32, 647–655. doi: 10.5555/3044805.3044879

Du, X., Ma, C., Zhang, G., Li, J., Lai, Y. K., Zhao, G., et al. (2020). An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.* 13, 1528–1540. doi: 10.1109/TAFFC.2020.3013711

Duan, L., Tsang, I. W., and Xu, D. (2012). Domain transfer multiple kernel learning. *IEEE Transac. Pattern Analy. Mach. Intellig.* 34, 465–479. doi: 10.1109/TPAMI.2011.114

Fang, C., Xu, Y., and Rockmore, D. N. (2013). "Unbiased metric learning: on the utilization of multiple datasets and web images for softening bias" in *Proceedings of IEEE International Conference on Computer Vision*. 1657–1664.

Feiping Nie, , Dong Xu, , Tsang, I. W. H., and Zhang, C. (2010). Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction. *IEEE Trans. Image Process.* 19, 1921–1932. doi: 10.1109/TIP.2010.2044958

Gan, C., Yang, T., and Gong, B. (2016). "Learning attributes equals multisource domain generalization" in *Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR)*. 87–97.

Ganin, Y., and Lempitsky, V. (2015). "Unsupervised domain adaptation by back-propagation" in Proceedings of International Conference of Machine Learning (ICML)., pp. 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030.

Gao, J., Huang, R., and Li, H. (2015). Sub-domain adaptation learning methodology (SDAL). *Inf. Sci.* 298, 237–256. doi: 10.1016/j.ins.2014.11.041

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2017). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Transac. Pattern Analy. Mach. Intellig.* 99:1. doi: 10.1109/TPAMI.2016.2599532

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. (2015). "Domain generalization for object recognition with multi-task autoencoders" in *Proceedings of IEEE International Conference of Computer Vision*. 2551–2559.

Gong, B., Shi, Y, Sha, F, and Grauman, K (2012). Geodesic flow kernel for unsupervised domain adaptation. Conference on Computer Vision and Pattern Recognition. IEEE. 2066–2073.

Gretton, A., Fukumizu, K, Harchaoui, Z, and Sriperumbudur, B.K. (2009). "A fast, consistent kernel two-sample test" in *Conference on Neural Information Processing Systems 22*, Vancouver, British Columbia, Canada. MIT Press. 673–681.

Han, Z., Chang, H., Zhou, X., Wang, J., Wang, L., and Shao, Y. (2022). E2ENNet: an end-to-end neural network for emotional brain-computer interface. *Front. Comput. Neurosci.* 16:942979. doi: 10.3389/fncom.2022.942979

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

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces. *IEEE Comput. Intell. Mag.* 11, 20–31. doi: 10.1109/MCI.2015.2501545

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/TAFFC.2014.2339834

Judy, H., Kulis, B., Darrell, T., and Saenko, K. (2012). Discovering latent domains for multisource domain adaptation. European Conference on Computer Vision. Springer, Berlin, Heidelberg.

Judy, H., Tzeng, E., Darrell, T., and Saenko, K. (2017). Simultaneous deep transfer across domains and tasks. *Domain Adaptat. Comput. Vision Appl.* 2017, 173–187. doi: 10.1007/978-3-319-58347-1_9

Kang, G., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. Contrastive adaptation network for single- and multi-source domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1793–1804. doi: 10.1109/TPAMI.2020.3029948

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. (2012). "Undoing the damage of dataset bias" in *Proceedings of European Conference on Computer Vision*. 158–171.

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Comput. Math. Methods Med.* 2013, 1–13. doi: 10.1155/2013/573734

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: a database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Muller-Putz, G. R. (2018). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Transac. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/TCDS.2018.2826840

Li, H., Jin, Y.-M., Zheng, W.-L., and Lu, B. L. (2018). "Cross-subject emotion recognition using deep adaptation networks" in *Neural Information Processing*. eds. L. Cheng, A. C. S. Leung and S. Ozawa (Cham: Springer International Publishing), 403–413.

Li, J., Qiu, S., Du, C., Wang, Y., and He, H. (2020). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Transac. Cogn. Dev. Syst.* 12, 344–353. doi: 10.1109/TCDS.2019.2949306

Li, J., Qiu, S., Shen, Y.-Y., Liu, C. L., and He, H. (2019). Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Transac. Cybernet.* 50, 3281–3293. doi: 10.1109/TCYB.2019.2904052

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* 12:162. doi: 10.3389/fnins.2018.00162

Li, R., Wang, Y., and Lu, B (2021). "A multi-domain adaptive graph convolutional network for EEG-based emotion recognition" in *Proceedings of the 29th ACM International Conference on Multimedia (MM'21)*, October 20–24. Virtual Event, China. ACM, New York, NY, USA, p. 9.

Li, W., Xu, Z., Xu, D., Dai, D., and van Gool, L. (2018). Domain generalization and adaptation using low-rank exemplar SVMs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1114–1127. doi: 10.1109/TPAMI.2017.2704624

Li, Y., Zheng, W., Cui, Z., Zhang, T., and Zong, Y. (2018b). "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition" in *The 27th International Joint Conference on Artificial Intelligence (IJCAI)*.

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018a). EEG emotion recognition based on graph regularized sparse linear regression. *Neural. Process. Lett.* 47, 1–19. doi: 10.1007/s11063-017-9609-3

Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2019). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 13, 568–578. doi: 10.1109/TAFFC.2019.2922912

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018c). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 12, 494–504. doi: 10.1109/TAFFC.2018.2885474

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks" in Proceedings of International Conference on Machine Learning. 97–105.

Long, M., Cao, Z., Wang, J., and Jordan, M. I.. (2018). "Conditional adversarial domain adaptation" in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 1647–1657.

Long, M., Wang, J., Ding, G., Pan, S. J., and Yu, P. S. (2014b). Adaptation regularization: a general framework for transfer learning. *IEEE Transac. Knowledge Data Eng.* 26, 1076–1089. doi: 10.1109/TKDE.2013.111

Long, M., Wang, J, Ding, G, Sun, J, and Philip, S.Y. (2014a). "Transfer joint matching for unsupervised domain adaptation" in *Conference on Computer Vision and Pattern Recognition*. IEEE. 1410–1417.

Long, M., Wang, J., and Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. *Proc. Int. Conf. Mach. Learn.* 70, 2208–2217. doi: 10.5555/3305890.3305909

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016). "Unsupervised domain adaptation with residual transfer networks" in *Proceedings of the Neural Information Processing Systems*. 136–144.

Luo, Y., Zhang, S. Y., Zheng, W. L., and Lu, B-L (2018). "WGAN domain adaptation for EEG-based emotion recognition" in International Conference on Neural Information Processing.

Mansour, Y, Mohri, M, and Rostamizadeh, A (2009). "Domain adaptation with multiple sources" in *Conference on Neural Information Processing Systems*. Vancouver, British Columbia, Canada, MIT Press. 1041–1048.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G (2017). "Unified deep supervised domain adaptation and generalization" in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy. pp. 5716–5726.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. *Proc. Int. Conf. Mach. Learn* 28, 10–18.

Mughal, N. E., Khan, M. J., Khalil, K., Javed, K., Sajid, H., Naseer, N., et al. (2022). EEG-fNIRS-based hybrid image construction and classification using CNN-LSTM. *Front. Neurorobot.* 16:873239. doi: 10.3389/fnbot.2022.873239

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interf.* 1, 66–84. doi: 10.1080/2326263X.2014.912881

Musha, T., Terasaki, Y., Haque, H. A., and Ivamitsky, G. A. (1997). Feature extraction from EEGs associated with emotions. *Artif. Life Robot.* 1, 15–19. doi: 10.1007/BF02471106

Nie, F, Huang, H, Cai, X, and Ding, C (2010). "Efficient and robust feature selection via joint-norms minimization" in *International Conference on Neural Information Processing Systems*. Curran Associates Inc. 1813–1821

Niu, L., Li, W., and Xu, D. (2015). "Visual recognition by learning from web data: a weakly supervised domain generalization approach" in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2774–2783.

Niu, L., Li, W., Xu, D., and Cai, J. (2018). An exemplar-based multi-view domain generalization framework for visual recognition. *IEEE Transac. Neural Netw. Learn. Syst.* 29, 259–272. doi: 10.1109/TNNLS.2016.2615469

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

Pandey, P., and Seeja, K. (2019). "Emotional state recognition with EEG signals using subject independent approach" in *Lecture Notes on Data Engineering and Communications Technologies, Data Science and Big Data Analytics* (Springer), 117–124.

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.* 32, 53–69. doi: 10.1109/MSP.2014.2347059

Pei, Z., Cao, Z., Long, M., and Wang, J. (2018). Multi-adversarial domain adaptation. *Proc. AAAI* 32, 3934–3941. doi: 10.1609/aaai.v32i1.11767

Rosenstein, M.T., Marx, Z., and Kaelbling, L. P. "To transfer or not to transfer" in *Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Shi, L. C., Jiao, Y. Y., and Lu, B. L. (2013). "Differential entropy feature for EEG-based vigilance estimation" in *The 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, *Osaka*. 6627–6630.

Simonyan, K., and Zisserman, A. (2014). "Very deep convolutional networks for large-scale image recognition" in *Proceedings of International Conference Learning Representations (ICLR)*. 1–14.

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). "EEG emotion recognition using dynamical graph convolutional neural networks" in *IEEE Transactions on Affective Computing*, pp. 1.

Sun, B., and Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. Proc. ECCV Workshops.

Tao, J., Chung, F. L., and Wang, S. (2012). On minimum distribution discrepancy support vector machine for domain adaptation. *Pattern Recogn.* 45, 3962–3984. doi: 10.1016/j.patcog.2012.04.014

Tao, J., and Dan, Y. (2021). Multi-source co-adaptation for EEG-based emotion recognition by mining correlation information. *Front. Neurosci.* 15:677106. doi: 10.3389/fnins.2021.677106

Tao, J., Dan, Y., Zhou, D., and He, S. (2022). Robust latent multi-source adaptation for encephalogram-based emotion recognition. *Front. Neurosci.* 16:850906. doi: 10.3389/fnins.2022.850906

Tao, J., Di, Z., Fangyu, L., and Bin, Z. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Pattern Recogn.* 87, 296–316. doi: 10.1016/j.patcog.2018.10.023

Tao, J., Song, D., Wen, S., and Hu, W. (2017). Robust multi-source adaptation visual classification using supervised low-rank representation. *Pattern Recogn.* 61, 47–65. doi: 10.1016/j.patcog.2016.07.006

Tommasi, T., Orabona, F., and Caputo, B. (2014). Learning Cate F.Ories from few examples with multi model knowledge transfer [J]. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 928–941. doi: 10.1109/TPAMI.2013.197

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA. 2962–2971.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K, and Darrell, T (2014). Deep domain confusion: Maximizing for domain invariance. arXiv [Preprint]. doi: 10.48550/arXiv.1412.3474

Wang, C., and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain. 1541–1546.

Wang, Y., Qiu, S., Li, D., Du, C., Lu, B. L., and He, H. (2022). Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition. *IEEE/CAA J. Automat. Sin.* 9, 1612–1626. doi: 10.1109/JAS.2022.105515

Xiao, M., and Guo, Y. (2015). Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE Transac. Pattern Analy. Mach. Intellig.* 37, 54–66. doi: 10.1109/TPAMI.2014.2343216

Xu, Y., He, X., Xu, G., Qi, G., Yu, K., Yin, L., et al. (2022). A medical image segmentation method based on multi-dimensional statistical features. *Front. Neurosci.* 16:1009581. doi: 10.3389/fnins.2022.1009581

Xu, Z., Li, W., Niu, L., and Xu, D. (2014). "Exploiting low-rank structure from latent domains for domain generalization" in *Proceedings of European Conference on Computer Vision*. 628–643

Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., and Lin, S. (2006). Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transac. Pattern Analy. Mach. Intellig.* 29, 40–51. doi: 10.1109/TPAMI.2007.250598

Yang, Y., Ma, Z., Hauptmann, A. G., and Sebe, N. (2013). Feature selection for multimedia analysis by sharing information among multiple tasks. *IEEE Trans. Multimed.* 15, 661–669. doi: 10.1109/TMM.2012.2237023

Yang, J., Yan, R., and Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive SVMs. *ACM Int. Conf. Multimedia ACM* 2007, 188–197. doi: 10.1145/1291233.1291276

Zhang, Y., Chen, J., Tan, J. H., Chen, Y., Chen, Y., Li, D., et al. (2020c). An investigation of deep learning models for EEG-based emotion recognition. *Front. Neurosci.* 14:622759. doi: 10.3389/fnins.2020.622759

Zhang, Y., Chung, F. L., and Wang, S. (2019a). Takagi-Sugeno-Kang fuzzy systems with dynamic rule weights. *J. Intell. Fuzzy Syst.* 37, 8535–8550. doi: 10.3233/JIFS-182561

Zhang, Y., Chung, F., and Wang, S. (2020a). Clustering by transmission learning from data density to label manifold with statistical diffusion. *Knowl.-Based Syst.* 193:105330. doi: 10.1016/j.knosys.2019.105330

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019b). Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition. *IEEE Access* 7, 127600–127614. doi: 10.1109/ACCESS.2019.2937657

Zhang, Y., Li, J., Zhou, X., Zhou, T., Zhang, M., Ren, J., et al. (2019c). A view-reduction based multi-view TSK fuzzy system and its application for textile color classification. *J. Ambient. Intell. Humaniz. Comput.* 2019, 1–11. doi: 10.1007/s12652-019-01495-9

Zhang, Y., Tian, F., Wu, H., Geng, X., Qian, D., Dong, J., et al. (2017). Brain MRI tissue classification based fuzzy clustering with competitive learning. *J. Med. Imag. Health Inform.* 7, 1654–1659. doi: 10.1166/jmihi.2017.2181

Zhang, Y., Wang, L., Wu, H., Geng, X., Yao, D., and Dong, J. (2016). A clustering method based on fast exemplar finding and its application on brain magnetic resonance images segmentation. *J. Med. Imag. Health Inform.* 6, 1337–1344. doi: 10.1166/jmihi.2016.1923

Zhang, Y., Wang, S., Xia, K., Jiang, Y., and Qian, P. (2020b). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Inform. Fusion* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zheng, W. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transac. Cogn. Dev. Syst.* 9, 281–290. doi: 10.1109/TCDS.2016.2587290

Zheng, W. L., and Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zheng, W. L., and Lu, B. L. (2016). "Personalizing EEG-based affective models with transfer learning" in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2732–2738.

Zheng, W. L., Zhang, Y. Q., Zhu, J. Y., and Lu, BL (2015). "Transfer components between subjects for EEG-based emotion recognition" in *International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an*. 917–922.

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* 13, 1290–1301. doi: 10.1109/TAFFC.2020.2994159

Zhong, Q., Zhu, Y., Cai, D., Xiao, L., and Zhang, H. (2020). Electroencephalogram Access for emotion recognition based on a deep hybrid network. *Front. Hum. Neurosci.* 14:589001. doi: 10.3389/fnhum.2020.589001

Zhou, R., Zhang, Z., Fu, H., Zhang, L, Li, L, Huang, G, et al. (2022). PR-PL: a novel transfer learning framework with prototypical representation based pairwise learning for EEG-based emotion recognition. arXiv [Preprint]. doi: 10.48550/arXiv.2202.06509

Zhu, Y., Zhuang, F., Wang, J., Ke, G., Chen, J., Bian, J., et al. (2020). Deep subdomain adaptation network for image classification. *IEEE Transac. Neural Netw. Learn. Syst.* 99, 1–10. doi: 10.1109/TNNLS.2020.2988928

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 09 November 2023
DOI 10.3389/fnins.2023.1247082

OPEN ACCESS

EDITED BY
Yue Zhao,
Harbin Institute of Technology, China

REVIEWED BY
Jianxing Liu,
Harbin Institute of Technology, China
Fangfang Duan,
Wuhan University of Technology, China

*CORRESPONDENCE
Di Zhou
✉ sion2005@sasu.edu.cn

†These authors have contributed equally to this work

RECEIVED 25 June 2023
ACCEPTED 20 October 2023
PUBLISHED 09 November 2023

CITATION
Tao J, Dan Y and Zhou D (2023) Possibilistic distribution distance metric: a robust domain adaptation learning method.
Front. Neurosci. 17:1247082.
doi: 10.3389/fnins.2023.1247082

Possibilistic distribution distance metric: a robust domain adaptation learning method

Jianwen Tao[1†], Yufang Dan[1†] and Di Zhou[2*]

[1]Institute of Artificial Intelligence Application, Ningbo Polytechnic, Zhejiang, China, [2]Industrial Technological Institute of Intelligent Manufacturing, Sichuan University of Arts and Science, Dazhou, China

The affective Brain-Computer Interface (aBCI) systems, which achieve predictions for individual subjects through training on multiple subjects, often cannot achieve satisfactory results due to the differences in Electroencephalogram (EEG) patterns between subjects. One tried to use Subject-specific classifiers, but there was a lack of sufficient labeled data. To solve this problem, Domain Adaptation (DA) has recently received widespread attention in the field of EEG-based emotion recognition. Domain adaptation (DA) learning aims to solve the problem of inconsistent distributions between training and test datasets and has received extensive attention. Most existing methods use Maximum Mean Discrepancy (MMD) or its variants to minimize the problem of domain distribution inconsistency. However, noisy data in the domain can lead to significant drift in domain means, which can affect the adaptability performance of learning methods based on MMD and its variants to some extent. Therefore, we propose a robust domain adaptation learning method with possibilistic distribution distance measure. Firstly, the traditional MMD criterion is transformed into a novel possibilistic clustering model to weaken the influence of noisy data, thereby constructing a robust possibilistic distribution distance metric (P-DDM) criterion. Then the robust effectiveness of domain distribution alignment is further improved by a fuzzy entropy regularization term. The proposed P-DDM is in theory proved which be an upper bound of the traditional distribution distance measure method MMD criterion under certain conditions. Therefore, minimizing P-DDM can effectively optimize the MMD objective. Secondly, based on the P-DDM criterion, a robust domain adaptation classifier based on P-DDM (C-PDDM) is proposed, which adopts the Laplacian matrix to preserve the geometric consistency of instances in the source domain and target domain for improving the label propagation performance. At the same time, by maximizing the use of source domain discriminative information to minimize domain discrimination error, the generalization performance of the learning model is further improved. Finally, a large number of experiments and analyses on multiple EEG datasets (i.e., SEED and SEED-IV) show that the proposed method has superior or comparable robustness performance (i.e., has increased by around 10%) in most cases.

KEYWORDS

electroencephalogram, domain adaptation, probabilistic clustering, maximum mean discrepancy, fuzzy entropy

1. Introduction

In the field of affective computing research (Mühl et al., 2014), automatic emotion recognition (AER) (Dolan, 2002) has received considerable attention from the computer vision community (Kim et al., 2013; Zhang et al., 2017). Thus far, numerous Electroencephalogram (EEG)-based emotion recognition methods have been proposed (Musha et al., 1997; Jenke et al., 2014; Zheng, 2017; Li X. et al., 2018; Pandey and Seeja, 2019). From a machine learning perspective, EEG-based AER can be modeled as a classification or regression problem (Kim et al., 2013; Zhang et al., 2017), where state-of-the-art AER techniques typically train their classifiers on multiple subjects to achieve accurate emotion recognition. In this case, subject-independent classifiers usually have poor generalization performance, as emotion patterns may vary across subjects (Pandey and Seeja, 2019). Significant progress in emotion recognition has been made by improving feature representation and learning models (Zheng et al., 2015; Zheng and Lu, 2015; Li et al., 2018a,b, 2019; Song et al., 2018; Du et al., 2020; Zhong et al., 2020). Since the individual differences in EEG-based AER are a natural existence, we may obtain a not good result by qualitative and empirical observations if the learned classifier generalize to previously unseen subjects (Jayaram et al., 2016; Zheng and Lu, 2016; Ghifary et al., 2017; Lan et al., 2019). As a possible solution, subject-specific classifiers are often impractical due to insufficient training data. Moreover, even if they are feasible in some specific scenarios, it is also an indispensable task to fine-tune the classifier to maintain a sound recognition capacity partly because the EEG signals of the same subject are changing now and then (Zhou et al., 2022). To address the aforementioned challenges, the domain adaptation (DA) learning paradigm (Patel et al., 2015; Tao et al., 2017, 2021, 2022; Zhang et al., 2019b; Dan et al., 2022) has been proposed and has achieved widespread effective applications, which enhances learning performance in the target domain by transferring and leveraging prior knowledge from other related but differently distributed domains (referred to as source or auxiliary domains), where the target domain has few or even no training samples.

Reducing or eliminating distribution differences between different domains is a crucial challenge currently faced during DA learning. To this end, mainstream DA learning methods primarily eliminate distribution biases between different domains by exploring domain-invariant features or samples (Pan and Yang, 2010; Patel et al., 2015). In order to fully exploit domain-invariant feature information, traditional shallow DA models have been extended to the deep DA paradigm. Benefiting from the advantages of deep feature transformation, deep DA methods have now achieved exciting adaptation learning performance (Long et al., 2015, 2016; Ding et al., 2018; Chen et al., 2019; Lee et al., 2019; Tang and Jia, 2019). Unfortunately, these deep DA methods can provide more transferable features and domain-invariant features, they can only alleviate but not eliminate the domain distribution shift problem caused by domain distribution differences. In addition, these deep DA methods can demonstrate better performance advantages, which may be attributed to one or several factors such as deep feature representation, model fine-tuning, adaptive regularization layers/terms, etc. However, the learning results of these methods still lack theoretical or practical interpretability at present.

DA theoretical studies have been proposed for domain adaptation generalization error bound (Ben-David et al., 2010) by the following inequality:

$$e_{\mathcal{T}}(h) \le e_{\mathcal{S}}(h) + d_{\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) + \min\left\{\varepsilon_{\mathcal{D}_{\mathcal{S}}}\left[\left|f_{\mathcal{S}}(x) - f_{\mathcal{T}}(x)\right|\right], \varepsilon_{\mathcal{D}_{\mathcal{T}}}\left[\left|f_{\mathcal{S}}(x) - f_{\mathcal{T}}(x)\right|\right]\right\}, \quad (1)$$

where the expected error of the target hypothesis $e_{\mathcal{T}}(h)$ is mainly constrained by three aspects: (1) the expected error of the source domain hypothesis $e_{\mathcal{S}}(h)$; (2) the distribution difference between the source and target domains $d_{\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$; (3) the difference in label functions between the two domains [i.e., the third term from Equation (1)]. Therefore, we will consider the three aspects simultaneously in this paper to reduce the domain adaptation generalization error bound (Zhang et al., 2021). Most existing methods assume that once the domain difference is minimized, a classifier trained only on the source domain can also generalize to the target domain well. Therefore, current mainstream DA methods aim to minimize the statistical distribution difference between the two domains. To this end, reducing or eliminating the distribution difference between domains to achieve knowledge transfer from the source domain and improve learning performance in the target domain is the core goal of domain adaptation learning methods. However, the key to this goal is effectively measuring the distribution difference between domains. Existing criteria for measuring the distance between different domains mainly include Maximum Mean Discrepancy (MMD) (Gretton et al., 2007), Bregman divergence, Jensen-Shannon divergence, etc. MMD is the most commonly used domain distribution difference measurement criterion in existing research, which can be divided into two categories alignment method: based on distribution alignment (including instance re-weighting and feature transformation) and classification model alignment with some representative works (Gretton et al., 2007; Pan et al., 2011; Tao et al., 2012, 2015, 2016, 2019; Baktashmotlagh et al., 2013; Chu et al., 2013; Long et al., 2013; Ganin et al., 2016; Liang et al., 2018; Luo et al., 2020; Kang et al., 2022).

To address the domain distribution shifting phenomenon, early instance re-weighting methods calculate the probability of each instance belonging to the source or target domain by likelihood ratio estimation (i.e., the membership of each instance). The domain shift problem can be relieved by re-weighting instances based on their membership. MMD (Gretton et al., 2007) is a widely adopted strategy for instance re-weighting, which is simple and effective. However, its optimization process is often carried out separately from the classifier training process, it's difficult to ensure that both are optimal at the same time. To address this issue, Chu et al. (2013) proposed a joint instance re-weighting DA classifier. To overcome the conditional distribution consistency assumption of the instance re-weighting method, the feature transformation methods have received widespread attention and exploration (Pan et al., 2011; Baktashmotlagh et al., 2013; Long et al., 2013; Liang et al., 2018; Luo et al., 2020; Kang et al., 2022). Representative methods include Pan et al. (2011) proposed the Transfer Component Analysis (TCA) method, which learned a transformation matrix. It adopted MMD technology to minimize the distribution distance between source domains and target domain, and preserved data divergence information, but did not consider domain

semantic realignment. Then, Long et al. (2013) proposed a Joint DA (JDA) method, which fully considered the domain feature distribution alignment and class conditional distribution alignment with the target domain labels in the class conditional distribution initialized by pseudo-labels. Recently, Luo et al. (2020) proposed a Discriminative and Geometry Aware Unsupervised Domain Adaptation (DGA-DA) framework, which combined the TCA and JDA methods. It introduced a strategy that made different classes from cross-domains mutually exclusive. Most of the existing affective models were based on deep transfer learning methods built with domain-adversarial neural network (DANN) (Ganin et al., 2016) proposed in Li et al. (2018c,d), Du et al. (2020), Luo et al. (2018), and Sun et al. (2022). The main idea of DANN (Ganin et al., 2016) was to find a shared feature representation for the source domain and the target domain with indistinguishable distribution differences. It also maintained the predictive ability of the estimated features on the source samples for a specific classification task. In addition, the framework preserved the geometric structure information of domain data to achieve effective propagation of target labels. Baktashmotlagh et al. (2013) proposed a Domain Invariant Projection (DIP) algorithm, which investigated the use of polynomial kernels in MMD to construct a compact domain-shared feature space. The series of DANN methods still has some challenges, PR-PL (Zhou et al., 2022) also explored the prototypical representations to further characterize the different emotion categories based on the DANN method. Finally, the study designed a clustering-based DA concept to minimize inner-class divergence. A review of existing DA method research shows that MMD is the main distribution distance measurement technique adopted by feature transformation-based DA methods. Traditional MMD-based DA methods focused solely on minimizing cross-domain distribution differences while ignoring the statistical (clustering) structure of the target domain distribution, which to some extent affects the inference of target domain labels. To address this issue, Kang et al. (2022) proposed a contrastive adaptation network based on unsupervised domain adaptation. The initialization of the labels from the target domain was realized by the clustering assumption. The feature representation is adjusted by measuring the contrastive domain differences (i.e., minimizing within-class domain differences and maximizing between-class domain differences) in multiple fully connected layers. During the training process, the assumptions of the target domain label and the feature representations are continuously cross-iterated and optimized to enhance the model's generalization capability. Furthermore, inspired by clustering patterns, Liang et al. (2018) proposed an effective domain-invariant projection integration method that uses clustering ideas to seek the best projection for each class within the domain, bridging the domain-invariant semantic gap and enhance the inner-class compactness in the domain. However, it still essentially belongs to MMD-based feature transformation DA methods.

It is worth noting that existing MMD-based methods did not fully consider the impact of intra-domain noise when measuring domain distribution distance. In real scenarios, noise inherently exists in domains, and intra-domain noise can lead to mean-shift problems in distance measurement for traditional MMD methods and their variants. This phenomenon to some extent is affecting the generalization performance of MMD-based DA methods. As shown in Figures 1A1, B1 represent the noise-free source domain and target domain, respectively. μ_{s*} and μ_{t*} are the means of the source domain and target domain, respectively. Figure 1C1 shows the domain adaptation result based on the MMD method. When the source domain has noises (i.e., Figure 1A2), the mean shift occurs and it's difficult to effectively measure the distribution distance by the MMD criterion. It matches the most of target domain samples (i.e., Figure 1B2) to a certain category of source domain (i.e., Figure 1C2). It declines the inferring performance of domain adaptation learning.

Existing research (Krishnapuram and Keller, 1993) pointed out that the possibilistic-based clustering model can effectively suppress noise interference during the clustering process. Therefore, Dan et al. (2021) proposed an effective classification model based on the possibilistic clustering assumption. Inspired by this work, we aim to jointly address the robustness and discriminative issues in the MMD criterion to enhance the adaptability of MMD-based methods and propose a robust Probabilistic Distribution Distance Measure (P-DDM) criterion. Specifically, by measuring the distance between EEG data (from either the source or target domain) and the overall domain mean (i.e., the mean of the source domain and target domain), the corresponding matching membership is used to judge the relevance between the EEG data and the mean. In other words, the smaller the distance between the EEG data and the mean, the larger the membership, and vice versa. In this way, the impact of noise in the matching process can be alleviated by the value of membership. The robustness and effectiveness of P-DDM are further enhanced by introducing a fuzzy entropy regularization term. Based on this, a domain adaptation Classifier model based on P-DDM (C-PDDM) is proposed, which introduces the graph Laplacian matrix to preserve the geometric structure consistency within the source domain and target domain. It can improve the label propagation performance. At the same time, a target domain classification model with better generalization performance is obtained by maximizing the use of source domain discriminative information to minimize domain discriminative errors. The main contributions of this paper are as follows:

1) The traditional MMD measurement is transformed into a clustering optimization problem, and a robust possibilistic distribution distance metric criterion (P-DDM) is proposed to solve the domain mean-shift problem in a noisy environment;
2) It is theoretically proven that under certain conditions, P-DDM is an upper bound of the traditional MMD measurement. The minimization of MMD in domain distribution measurement can be effectively achieved by optimizing the P-DDM;
3) A DA classifier mode based on P-DDM is proposed (i.e., C-PDDM), its consistent convergence is proven, and the DA generalization error bound of the method is proposed based on Rademacher complexity theory;
4) A large number of experiments are conducted on two EEG datasets (i.e., SEED and SEED-IV), demonstrating the robust effectiveness of the method and a certain degree of improvement in the classification accuracy of the model.

2. Proposed framework: C-PDDM

In domain adaptation learning, $\mathcal{D}_S = \left\{x_i^s, y_i^s\right\}_{i=1}^n$ denotes n samples and its associated labels of the source domain. $X^s = \left[x_1^s, \ldots, x_n^s\right] \in \mathbb{R}^{d\times n}$ indicates all the source samples.

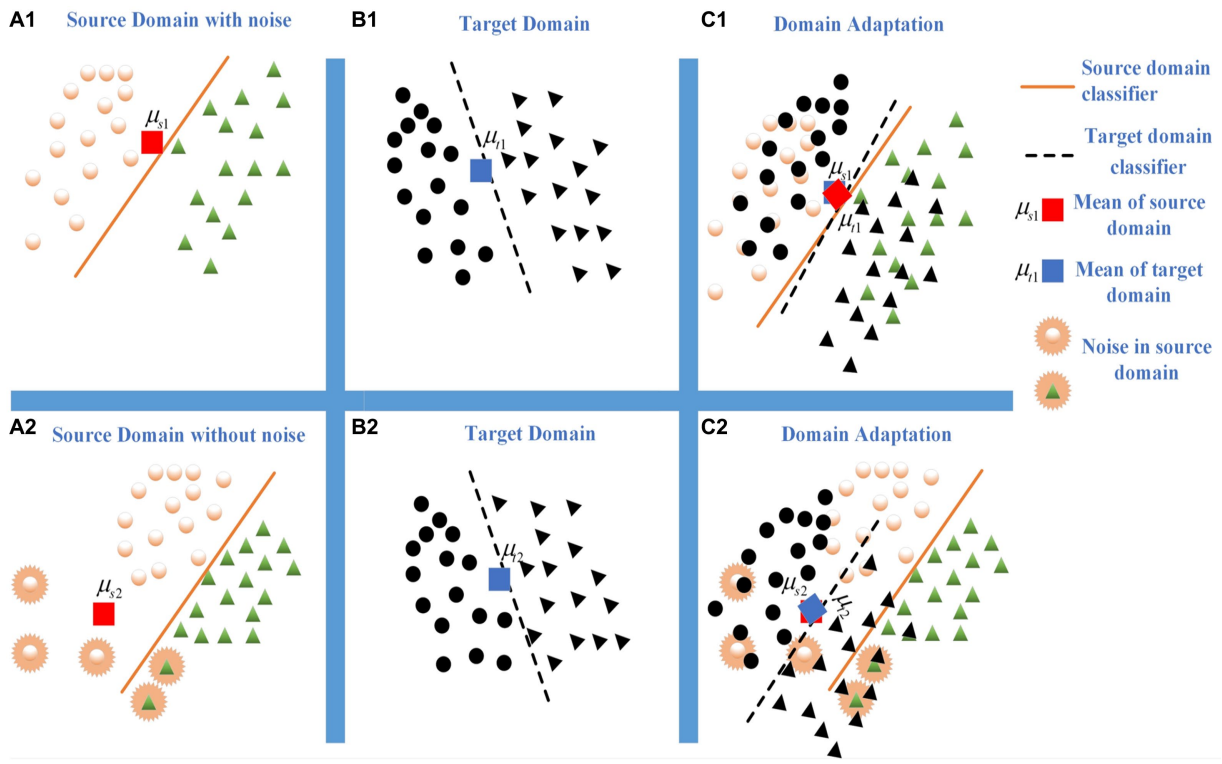


FIGURE 1
The influence comes from the noises or outliers during domain matching. **(A1)** Source domain with noise. **(B1)** Target domain. **(C1)** Domain adaptation. **(A2)** Source domain without noise. **(B2)** Target domain. **(C2)** Domain adaptation.

$Y^s = \{y_1,\ldots,y_n\}^T \in \{0,1\}^{n\times C}$ is the associated labels with a one-hot coding vector $y_i \in \mathbb{R}^C (1 \le i \le n)$. If x_i belongs to the j-*th* class, The other elements y_i are zero. $\mathcal{D}_T = \{x_j^t\}_{j=1}^m$ denotes the target domain with no label, which $X^t = [x_1^t,\ldots,x_m^t] \in \mathbb{R}^{d\times m}$ means m data points. $Y^t = \{y_1,\ldots,y_m\}^T \in \mathbb{R}^{m\times C}$ is unknown during training. Let $X = [X^s, X^t] \in \mathbb{R}^{d\times N}$, $N = n + m$, $\mu_s = \frac{1}{n}\sum_{i=1}^{n} x_i^s$, and $\mu_t = \frac{1}{m}\sum_{j=1}^{m} x_j^t$ denotes the mean value of the source domain and target domain, respectively. Our proposal has some assumptions:

1) However, the distributions of source domain ($\mathbb{P}$) and target domain ($\mathbb{Q}$) are different (i.e., $\mathbb{P}(\mathcal{X}_S) \ne \mathbb{Q}(\mathcal{X}_T)$ and $\mathcal{X}_S = \mathcal{X}_T$), they share the same feature space with $\mathcal{X}_S, \mathcal{X}_T \in \mathcal{X}$ are feature space of the source domain and target domain, respectively.
2) The condition probability distributions of the source domain and target domain are different [i.e., $\mathbb{P}(\mathcal{Y}_S|\mathcal{X}_S) \ne \mathbb{Q}(\mathcal{Y}_T|\mathcal{X}_T)$], but they share the same label space with $\mathcal{Y}_S, \mathcal{Y}_T \in \mathcal{Y}$ are label space of the source domain and target domain, respectively.

In the face of a complex and noisy DA environment, the proposed method will achieve the following objectives by the DA generalization error theory (Ben-David et al., 2010) to make the distance metric for domain adaptation more robust and achieve good target classification performance: (1) Robust distance metric: solve the problem of domain mean shift under the influence of noise, thereby effectively aligning the domain distribution differences; (2) Implement target domain knowledge inference: we bridge the discriminative information of the source domain while minimizing the domain discriminative error based on preserving the consistency of domain data geometry, and learn a target domain classification machine with high generalization performance. Based on the descriptions of the above objectives, the general form of the proposed method can be described as:

$$\Theta(\lambda_i, Y, W) = \min \Omega(\lambda_k, X^s, X^t) + R(Y, W) \qquad (2)$$

where $\Omega(\lambda_k, X^s, X^t)$ is the robust distance metric, which reduces the impact of noisy data on the alignment of domain distribution differences. *R(Y, W)* is the domain adaptation learning loss function that includes the label matrix Y (that is, the comprehensive label matrix of the source and target domains) and the comprehensive learning model W of the source domain and the target domain.

2.1. Design of possibilistic distribution distance metric

2.1.1. Motivation

In a certain reproducing kernel Hilbert space (RKHS)$\mathcal{H}$, the original space data representation can be transformed into a feature representation in the RKHS through a certain non-linear transformation $\phi: \mathbb{R}^d \to H$ (Long et al., 2016). The corresponding kernel function is defined as $K(X_1, X_2): X \times X \to \mathbb{R}$, where $K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$, $x_1, x_2 \in X$. It is also a commonly used kernel technique in current non-linear learning methods (Pan et al.,

2011; Long et al., 2015). For the problem of inconsistent distributions in domain adaptation, existing research has shown (Bruzzone and Marconcini, 2010; Gretton et al., 2010) that when sample data is mapped to a high-dimensional or even infinite-dimensional space, it can capture higher-dimensional feature representations of the data (Carlucci et al., 2017). That is, in a certain RKHS, the distance between two distributions can be effectively measured through the maximum mean discrepancy (MMD) criterion. Based on this, it is assumed that $\mathcal{F}$ is a collection of functions of a certain type $f: f: \mathcal{X} \to \mathbb{R}$, The maximum mean discrepancy (MMD) between two domain distributions $\mathbb{P}$ and $\mathbb{Q}$ can be defined as:

$$MMD_{\mathcal{F}}[\mathbb{P},\mathbb{Q}] := \sup_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathbb{P}}[f(x)] - \mathop{\mathbb{E}}_{\mathbb{Q}}[f(x)]\right). \tag{3}$$

MMD measure minimizes the expected difference between two domain distributions through the function *f*, making the two domain distributions as similar as possible. When the sample size of the domain is sufficiently large (or approaches infinity), the expected difference approximates (or equals) the empirical mean difference. Therefore, Equation (3) can be written in the empirical form of MMD:

$$MMD\left(X^s, X^t\right) := \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i^s\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(x_j^t\right)\right\|_H^2. \tag{4}$$

To prove the universal connection between the traditional MMD criterion and the mean clustering model, we give the following theorem: **Theorem 1**. The MMD measure can be loosely modeled as a special clustering problem with one cluster center, where the clustering center is μ, and the instance clustering membership is ς_k.

Proof: As defined by MMD:

$$\begin{aligned}
&MMD\left(D^s, D^t\right)\\
&= \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i^s\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(x_j^t\right)\right\|_H^2\\
&= \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i^s\right) - \mu + \mu - \frac{1}{m}\sum_{j=1}^{m}\phi\left(x_j^t\right)\right\|_H^2\\
&\le \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i^s\right) - \mu\right\| + \left\| \frac{1}{m}\sum_{j=1}^{m}\phi\left(x_j^t\right) - \mu\right\|_H^2\\
&= \frac{1}{n^2}\left\| \sum_{i=1}^{n}\phi\left(x_i^s\right) - n\mu\right\| + \frac{1}{m^2}\left\| \sum_{j=1}^{m}\phi\left(x_j^t\right) - m\mu\right\|_H^2\\
&= \frac{1}{n^2}\left\| \sum_{i=1}^{n}\left(\phi\left(x_i^s\right) - \mu\right)\right\|_H^2 + \frac{1}{m^2}\left\| \sum_{j=1}^{m}\left(\phi\left(x_j^t\right) - \mu\right)\right\|_H^2\\
&\le \frac{1}{n^2}\sum_{i=1}^{n}\left\|\phi\left(x_i^s\right) - \mu\right\|_H^2 + \frac{1}{m^2}\sum_{j=1}^{m}\left\|\phi\left(x_j^t\right) - \mu\right\|_H^2\\
&= \sum_{k=1}^{N}\varsigma_k\left\|\phi\left(x_k\right) - \mu\right\|_H^2
\end{aligned} \tag{5}$$

where $\mu = \delta\mu_s + (1-\delta)\mu_t$ is the cluster center with $0 \le \delta \le 1$. When $n = m$, let $\delta = 0.5$. When $n \ne m$, the number of samples in the source domain and target domain can be set the same during sampling. The sample membership ς_k of one cluster center is defined as:

$$\varsigma_k = \begin{cases} \dfrac{1}{n^2}, & x_k \in X^s \\ \dfrac{1}{m^2}, & x_k \in X^t \end{cases}. \tag{6}$$

From Equation (5), it can be seen that the one cluster center form with clustering center n is an upper bound of the traditional MMD measure. In other words, the MMD measure can be relaxed to a special one cluster center objective function. By optimizing this clustering objective, the minimization of MMD between domains can be achieved.

As indicated in Theorem 1 and Baktashmotlagh et al. (2013), the domain distribution MMD criterion is essentially related to the clustering model, which can be used to achieve more effective distribution alignment between different domains by clustering domain data. It is worth noting that the traditional clustering model has the disadvantage of being sensitive to noise (Krishnapuram and Keller, 1993), which makes domain adaptation (DA) methods based on MMD generally face the problem of domain mean shift caused by noisy data. To address this issue, this paper further explores more robust forms of clustering and proposes an effective new criterion for domain distribution distance measurement.

2.1.2. P-DDM

Recently proposed possibility clustering models can effectively overcome the impact of noise on clustering performance (Dan et al., 2021). Therefore, this paper further generalizes the above special one cluster center to a possibility one cluster center form and proposes a robust possibility distribution distance metric criterion P-DDM. By introducing the possibility clustering assumption, the MMD hard clustering form is generalized to a soft clustering form, which controls the contribution of each instance according to its distance from the overall domain mean. The farther the distance, the smaller the contribution of the instance, thus weakening the influence of mean shift caused by noisy data in the domain and improving the robustness of domain adaptation learning.

To achieve robust domain distribution alignment, the distribution distance measurement criterion based on the possibility clustering assumption mainly achieves two goals: (1) Calculate the difference in distribution between kernel space domains based on the possibility clustering assumption, by measuring the distance between each instance in the domain and the overall domain mean; (2) Measure the matching contribution of each instance. Any instance in the overall domain has a matching contribution value $\lambda_k \in \mathbb{R}$, $k = 1, 2, \ldots N$, which is the matching contribution degree of x_k to the overall domain mean, and the closer the distance, the larger the value of λ_k. Thus, the possibility distribution distance measure can be defined as:

$$\begin{aligned}
&\Omega_p\left(\lambda_k, X_s, X_t\right) = \sum_{k=1}^{N}\lambda_k^b\left\|\phi\left(x_k\right) - \mu\right\|_H^2,\\
&s.t., 0 \le \lambda_k \le 1, k = 1, \ldots, N
\end{aligned} \tag{7}$$

where the parameter b is the weight exponent of λ_k, which is used to adjust the uncertainty or degree of the data points belonging to

multiple categories. In order to circumvent the trivial solution, b is set to 2 in the subsequent equations of this paper. The detailed process of different values of b can be found in references (Krishnapuram and Keller, 1993). $\Omega_p(\lambda_k, X_s, X_t)$ is an objective function of possibility clustering with a cluster center of μ, and when $\lambda_k^2 = \varsigma_k$, $\Omega_p(\lambda_k, X_s, X_t)$ takes the form of the above-mentioned special one cluster center.Theorem 2.When $\lambda_k \in \left[\frac{1}{r}, 1\right]$, the possibility distribution distance measure $\Omega_p(\lambda_k, X_s, X_t)$ is an upper bound of the traditional MMD method.

Proof: Combining Equation (5) and Equation (7), we have the following inference process:

$$\begin{aligned} &\min_K MMD\left(X^s, X^t\right) \\ &\le \sum_{k=1}^{N} \varsigma_k \left\|\phi(x_k) - \mu\right\|_H^2 \\ &\le \sum_{k=1}^{N} \lambda_k^2 \left\|\phi(x_k) - \mu\right\|_H^2 \\ &= \Omega_p(\lambda_k, X_s, X_t) \end{aligned} \tag{8}$$

According to the value range of ς_k, when $\lambda_k \in \left[\left(1/r\right), 1\right]$ and $r = min(n, m)$, the second inequality in Equation (8) holds, thus proving that $\Omega_p(\lambda_k, X_s, X_t)$ is the upper bound of traditional MMD. According to Theorem 1 and Theorem 2, the traditional MMD metric criterion can be modeled as a possibilistic one cluster center objective form. From this perspective, it can be considered that the possibilistic distribution distance metric target domain can not only achieve alignment of domain feature distribution, but also weaken the "negative transfer" effect of noisy data in the domains during training.

Equation (7) only considers the overall mean regression problem, which clusters each instance with the overall domain mean, while ignoring the semantic structural information of the instance in domain distribution alignment. It may lead to the destruction of the local class structure in the domain. Inspired by the idea of global and local from Tao et al. (2016), we further consider the semantic distribution structure in domain alignment and calculate the semantic matching contribution of each instance. Therefore, based on the feature distribution alignment, we propose an integrated semantic alignment. It can be rewritten as follows:

$$\begin{aligned} &\Omega_{pc}(\lambda_k, X_s, X_t) \\ &= \min_{\lambda_{k,c}} \sum_{c=0}^{C} \sum_{k=1}^{N} \lambda_{k,c}^2 \left\|\phi(x_{k,c}) - \mu_c\right\|_H^2 \\ &s.t., 0 \le \lambda_{k,c} \le 1 \end{aligned} \tag{9}$$

where $\mu_c = \delta\mu_{s,c} + (1-\delta)\mu_{t,c}$, $\mu_{s,c} = \frac{1}{n}\sum_{c=0}^{C}\sum_{i=1}^{n_c}\phi\left(x_{i,c}^s\right)$, $\mu_{t,c} = \frac{1}{m}\sum_{c=0}^{C}\sum_{j=1}^{m_c}\phi\left(x_{j,c}^t\right)$, $c = 0,1,2,\ldots,C$, C is the number of classes. n_c is the number of samples of the c-th class in the source domain, m_c is the sample number of the c-th class in the target domain, and $n = \sum_{c=0}^{C} n_c$, $m = \sum_{c=0}^{C} m_c$. When $m = \sum_{c=0}^{C} m_c$. When $c = 0$, $\mu_{s,c}$ and $\mu_{t,c}$ are the mean values of the source domain and the target domain, respectively. Equation (9) is a feature distribution alignment form. When $c \in [1,2,\ldots,C]$, $\mu_{s,c}$ and $\mu_{t,c}$ are the associated c-th class mean values of the source domain and the target domain, respectively. $\lambda_{k,c}$ is the membership of x_k belonging to the c-th class in the overall domain (i.e., integrate the source domain and target domain into one domain).

To further improve the robustness and effectiveness of the possibilistic distribution distance metric method on noisy data, we add a fuzzy entropy regularization term related in Equation (9). Therefore, the semantic alignment P-DDM in (9) can be further defined as follows:

$$\begin{aligned} &\Omega(\lambda_k, X_s, X_t) \\ &= \sum_{c=0}^{C}\sum_{k=1}^{N} \lambda_{k,c}^2 \left\|\phi(x_{k,c}) - \mu_c\right\|_H^2, \\ &+ \beta \sum_{c=0}^{C}\sum_{k=1}^{N} \left(\lambda_{k,c}^2 \ln \lambda_{k,c}^2 - \lambda_{k,c}^2\right) \\ &s.t., 0 \le \lambda_{k,c} \le 1 \end{aligned} \tag{10}$$

where β is a tunable balancing parameter that forces the value of $\lambda_{k,c}$ for relevant data to be as large as possible to avoid trivial solutions. After the above improvements, P-DDM is a monotonic decreasing function on $\lambda_{k,c}$. Through the fuzzy entropy term in the second part of Equation (10), P-DDM reduces the impact of noise data on model classification. The larger the fuzzy entropy, the greater the sample discrimination information, which helps to enhance the robustness and effectiveness of distribution distance measurement. Additionally, the possibility distribution measurement model regularized by fuzzy entropy can effectively suppress the contribution of noise data in domain distribution alignment, thereby reducing the interference of noise/abnormal data to domain adaptation learning. The robustness effect of fuzzy entropy can be further seen in the empirical analysis of reference (Gretton et al., 2010).

2.2. Design of domain adaptation function

The P-DDM criterion addresses the problems of domain distribution alignment and noise impact. Next, we will achieve the two goals required for the inference of target domain knowledge: (1) to preserve the geometric consistency in the source domain and the target domain, i.e., the label information between adjacent samples should be consistent, and (2) to minimize the structural risk loss of both the source and target domains. Given the description of the objective task, the general form of the objective risk function can be described as:

$$R(Y, W) = \Omega_Y + \Omega_W, \tag{11}$$

where Ω_Y is the loss of joint knowledge transfer and label propagation, which preserves the geometric consistency of the data between the source and target domains, and Ω_W is the structural risk loss term, which includes both the source domain and the target domain. Next, these two terms will be designed separately.

2.2.1. Joint knowledge transfer and label propagation

Firstly, $G=\langle X,M\rangle$ denotes an undirected weighted graph of the overall domain. $M\in\mathbb{R}^{N\times N}$ is a weighted matrix with $M_{ij}=M_{ji}\geq 0$. M_{ij} is calculated by:

$$M_{ij}=\begin{cases}\exp\left(-\dfrac{\|x_i-x_j\|^2}{2\sigma^2}\right), x_i\in Ne(x_j)\text{或}x_j\in Ne(x_i),\\ 0,\text{otherwise}\end{cases} \tag{12}$$

where $x_k\in Ne(x_m)$ means that x_k is the neighbor of x_m. σ is the local influence range parameter that controls the Gaussian kernel function and is also a hyper-parameter. The larger the value of σ, the larger the local influence range, and vice versa, the smaller the local influence range. When σ is fixed, the change in M_{ij} decreases monotonically as the distance between x_i and x_j increases.

In combination with source domain knowledge transfer and graph Laplacian matrix (Long et al., 2013; Wang et al., 2017), the objective form of label propagation modeling can be described as:

$$\Omega_Y=\min_Y \mathrm{tr}\left(\mathrm{Y}^T\mathrm{LY}\right), \tag{13}$$

where $\mathrm{Y}=\left[\mathrm{Y}_s;\mathrm{Y}_t\right]\in\mathbb{R}^{N\times C}$, Y_t is the target domain label matrix. The label value for a sample in the target domain corresponding to a position in Y_t is all zeros when the sample has no label. Y_s is the source domain label matrix. $L=M-D\in\mathbb{R}^{N\times N}$ is the Laplacian graph matrix (Long et al., 2013) with D is a diagonal matrix and $D_{ii}=\sum_{j=1}^{N}M_{ij}$.

2.2.2. Minimize structural risk loss

In our proposed method, the classifier of the source domain (the corresponding target domain classification model) is defined as $f_s=W_{ss}{}^TX_s+b_s$ (the corresponding $f_t=W_{tt}{}^TX_t+b_t$). $b_s(b_t)$ is the source domain bias (the target source bias). $W_{ss}(W_{tt})$ is the parameter matrix of the source domain (the parameter matrix of the target domain). Let $\widetilde{W_s}=[W_{ss},b_s]$, $\widetilde{X_s}=[X_s,1]$, $\widetilde{W_t}=[W_{tt},b_t]$, $\widetilde{X_t}=[X_t,1]$, we can rewrite both classifiers of the source domain and the target domain respectively: $\widetilde{f_s}=\widetilde{W_s}^T\widetilde{X_s}$ and $\widetilde{f_t}=\widetilde{W_t}^T\widetilde{X_t}$. Let $W=\left[\widetilde{W_s},\widetilde{W_t}\right]$, $X=\left[\widetilde{X_s},\widetilde{X_t}\right]$. We rewrite the final classifier as: $F(W)=X^TW$.

According to the minimum square loss function, the problem of minimizing structural risk loss in both domains (source domain and target domain) can be described as:

$$\Omega_W=\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T(x_k)W-y_k\right\|_H^2+\rho\|W\|_{2,1} \tag{14}$$

where the first term denotes the structure risk loss and $y_k\in Y$ The second term is the constraint term of W. By using $l_{2,1}$ regularization, we can achieve feature selection and it can effectively control the complexity of the model to prevent over-fitting of the target classification model to some extent.

The classification task proposed in this method is ensured by the dual prediction of the label matrix Y and the decision function W to guarantee the reliability of the prediction. The target classification function is combined by Equation (13) and Equation (14). It's described as follows:

$$\begin{aligned}&R(Y,W)=\alpha tr\left(Y^TLY\right)+\\&\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T(x_k)W-y_k\right\|_H^2+\rho\|W\|_{2,1},\\&s.t.,0\leq\lambda_{k,c}\leq 1,YY^T=I\end{aligned} \tag{15}$$

2.3. Final formulation

By combining the semantic alignment P-DDM form [i.e., Equation (10)] and the target classification function [i.e., Equation (16)], the final optimization problem formulation of the proposed method C-PDDM can be described as follows:

$$\begin{aligned}&\Theta(\lambda_k,Y,W)\\&=\min_{\lambda_k,Y,W}\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi(x_{k,c})-\mu_c\right\|_H^2\\&+\beta\sum_{c=0}^{C}\sum_{k=1}^{N}\left(\lambda_{k,c}^2\ln\lambda_{k,c}^2-\lambda_{k,c}^2\right)+\alpha tr\left(Y^TLY\right),\\&+\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T(x_k)W-y_k\right\|_H^2+\rho\|W\|_{2,1},\\&s.t.,0\leq\lambda_{k,c}\leq 1,YY^T=I\end{aligned} \tag{16}$$

where β,α, and ρ are balance parameters.

With all model parameters obtained, target domain knowledge inference can be achieved by maximizing the utilization of source domain discriminative information, linearly fusing the two classifiers $\widetilde{f_s}$ and $\widetilde{f_t}$, and using this linear fusion model for target domain knowledge inference. The fusion form can be written as follows:

$$j=\arg\max_j\left(y_i^t=\upsilon\widetilde{f_s}(x_i^t)+(1-\upsilon)\widetilde{f_t}(x_i^t)\right)_j$$

where $\upsilon\in[0,1]$ is an adjustable parameter that balances the two classifiers, in order to reflect the importance of source domain discriminative information as prior knowledge, υ is set to 0.9 based on empirical experience.

3. C-PDDM optimization

The optimization problem of C-PDDM is a non-convex problem with respect to $\lambda_{k,c}$, W, and Y. We will adopt an alternating iterative optimization strategy to achieve the optimization and solution of $\lambda_{k,c}$, W, and Y, so that each optimization variable has a closed-form solution.

3.1. Update $\lambda_{k,c}$ as given W and Y

As we fix W and Y, the objective function in Equation (16) reduces to solving:

$$\begin{aligned}\min_{\lambda_{k,c}} P_1 &= \min \sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi\left(x_{i,c}\right)-\mu_c\right\|_H^2 \\ &-\beta\sum_{c=0}^{C}\sum_{k=1}^{N}(-\lambda_{k,c}^2\ln\lambda_{k,c}^2+\lambda_{k,c}^2)+ \\ &+\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T(x_i)W-y_i\right\|_H^2 \\ s.t. 0 &\le \lambda_{k,c}\le 1\end{aligned} \tag{17}$$

Theorem 3. The optimal solution to the primal optimization problem of the objective function (17) is:

$$\lambda_{k,c}=\exp\left(-\frac{J}{\beta}\right), \tag{18}$$

where $J=\sum_{c=0}^{C}\sum_{k=1}^{N}\left\|\phi^T\left(x_k\right)W-y_k\right\|_H^2+\sum_{c=0}^{C}\sum_{k=1}^{N}\left\|\phi\left(x_{k,c}\right)-\mu_c\right\|_H^2$.

Proof. By setting the derivative $\partial P_1/\partial\lambda_{k,c}=0$, we obtain:

$$\begin{aligned}\frac{\partial P_1}{\partial\lambda_{k,c}} &= 2\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}\left\|\phi\left(x_{k,c}\right)-\mu_c\right\|_H^2 \\ &+2\beta\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}\ln\lambda_{k,c}^2 \\ &+2\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}\left\|\phi^T\left(x_k\right)W-Y\right\|_H^2=0\end{aligned} \tag{19}$$

Combining and simplifying the terms in Equation (19), we get the solution of $\lambda_{k,c}$ is Equation (18), Theorem 3 is proved. From Theorem 3, the membership of any sample can be obtained by Equation (18).

3.2. Update W as given Y and $\lambda_{k,c}$

Since the first and the third terms in Equation (16) do not have W, the optimization formula for C-PDDM can be rewritten as:

$$\begin{aligned}P_2 &= \min_W \sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T\left(x_k\right)W-y_k\right\|_H^2+\rho\|W\|_{2,1} \\ &= \min_W \lambda\left\|\phi^T\left(X\right)W-Y\right\|_H^2+\rho\|W\|_{2,1}\end{aligned}, \tag{20}$$

where λ is a matrix with $\lambda\in\mathbb{R}^{N\times C}$, each element is $\lambda_{k,c}^2$, $\lambda_{k,c}$ means the membership of x_k belonging to the *c-th* class.Theorem 4.The optimal solution to the primal optimization problem of the objective function (20) is:

$$\mathrm{W}=\mathrm{AY}, \tag{21}$$

with $A=\left(\lambda\phi(\mathrm{X})\phi^T(\mathrm{X})+\rho U\right)^{-1}\phi(\mathrm{X})$.

Proof. According to Equation (19), let $\partial P_2/\partial W=0$, we have:

$$\begin{aligned}&\partial P_2/\partial W \\ &=2\lambda\left[\phi(X)\left(\phi^T(X)\mathrm{W}-Y\right)\right]+2\rho UW, \\ &=0\end{aligned} \tag{22}$$

where $\frac{\partial\rho\|W\|_{2,1}}{\partial W}=UW$, U is a diagonal matrix, its diagonal element is $U_{ii}=\frac{1}{\|w_i\|}$, w_i is the *i*-th vector of W. The solution obtained by organizing Equation (22) is Equation (21).

3.3. Update Y by fixing W and $\lambda_{k,c}$

Finally, $\lambda_{k,c}$ is fixed. W = AY is substituted into Equation (16). The constraint $\mathrm{YY}^T=I$ can reduce the interference information in the label matrix Y, the objective form for optimizing the solution of Y is described as:

$$\begin{aligned}P_3 &= \min_{Y^TY=I}\alpha tr(Y^TLY) \\ &+\sum_{c=0}^{C}\sum_{k=1}^{N}\lambda_{k,c}^2\left\|\phi^T(x_k)W-y_k\right\|_H^2 \\ &= \min_{Y^TY=I}\alpha tr(Y^TLY) \\ &+\lambda\left\|\phi^T(X)AY-Y\right\|_H^2 \\ &= \min_{Y^TY=I} tr\left(Y^THY\right)\end{aligned} \tag{23}$$

where $H=\alpha\mathrm{L}+\lambda B^TB, B=\phi^T\left(X\right)\mathrm{A}-I$.

The optimization problem (23) is a standard singular value decomposition problem, where Y is the eigenvector of the matrix H. Y can be obtained by solving the singular value decomposition of the matrix H.

4. Algorithm

4.1. Algorithm description

In unsupervised domain adaptation learning scenarios (i.e., the target domain does not have any labeled data), in order to achieve semantic alignment between domains, initial labels of the target domain can be obtained through three strategies (Liang et al., 2018): (1) random initialization; (2) zero initialization; (3) use the model trained on the source domain data to cluster the target domain data to obtain initial labels. (1) and (2) belong to the cold-start method. (3) belongs to the hot-start method which is relatively friendly to subsequent learning performance. Therefore, we adopt the third method to initialize the prior information of $\lambda_{k,c}$, W, and Y. The proposed method adopts the iterative optimization strategy commonly used in multi-objective optimization, and the algorithm stops iterating when the following conditions are satisfied: $\left|\Theta\left(\lambda_{k,c}^z,W^z,Y^z\right)-\Theta(\lambda_{k,c}^{z-1},,W^{z-1},,Y^{z-1})\right|<\varepsilon$, where $\Theta\left(\lambda_{k,c}^z,W^z,Y^z\right)$ denotes the value of the objective function at the *z-th* iteration. ε is a pre-defined threshold.

ALGORITHM 1 Domain adaptation learning based on C-PDDM.

Input: The source domain data $\{X_S, Y_S\}$, the target domain data X_t, unknown labels of the target domain Y_t (the initialization can be obtained by cluster algorithm), model parameter values of $\beta, \alpha, \rho, \theta$ and the threshold of iteration stop ε, and the maximal iteration number Z.
Output: The contribution matrix $\lambda_{k,c}$ matches each instance to the mean points of each class in the entire domain, the decision function W and the label matrix Y.
Procedure:
1. Initialize the label values for unlabeled data from the target domain.
2. Compute the means of different classes in the target domain and the source domain respectively, denoted as $\mu_{t,c}$ and $\mu_{s,c}$, c = 0,1,2,…,C.
3. Then compute the mean of different class data in the overall domain (i.e., integrate the source domain and the target domain), denoted as $\mu_c = \frac{1}{2}(\mu_{s,c} + \mu_{t,c})$
4. Obtain the initialization $\lambda_{k,c}^0$ of $\lambda_{k,c}$ using (18);
5. Obtain the initialization W^0 of W using (21);
6. Obtain the initialization Y^0 of Y using (23);
7. Compute the value of the objective function $\tilde{} \left(\lambda_{k,c}^0, W^0, Y^0\right)$;
8.**for** $z = 1$ **to** Z **do**:
{
8.1 Fix the current W and Y for updating $\lambda_{k,c}$ to $\lambda_{k,c}^z$ by Eq. (18) ;
8.2 Fix the current $\lambda_{k,c}$ and Y for updating W to W^z by Eq. (21) ;
8.3 Fix the current $\lambda_{k,c}$ and W for updating Y to Y^z by Eq. (23) ;
}
while $\left| \tilde{} \left(\lambda_{k,c}^z, W^z, Y^z\right) - \tilde{} \left(\lambda_{k,c}^{z-1}, W^{z-1}, Y^{z-1}\right) \right| \geq \varepsilon$
9. return $\lambda_{k,c}$, W, and Y;

4.2 Computational complexity

This article uses Big O to analyze the computational complexity of Algorithm 1. The proposed method C-PDDM mainly consists of two joint optimization parts: P-DDM and target label propagation. Specifically, we first construct the *k-Nearest Neighbor (*i.e., *k-NN*) graph and compute the kernel matrix K in advance requiring computational costs of $O\left(dn^2\right)$ and $O\left(dN^2\right)$, respectively. Then, the optimization process of Algorithm 1 requires T iterations to complete with the P-DDM minimization (including possibility membership inference) process requires $O\left(d^3 + N^2 + d^2N\right)$. The target label matrix F_t requires $O\left(3n^3 + n^2c\right)$ to complete inferring thing. The target classification model W requires $O\left(nc^2 + dc^2\right)$ to finish updating, Therefore, the overall computational cost of Algorithm 1 is $O\left(T\left(d^3 + N^2 + d^2N + 3n^3 + n^2c\right) + dn^2 + dN^2\right)$.

Before training in Algorithm 1, pre-computing the C-PDDM kernel matrix and Laplacian graph matrix and loading them into memory can further improve the computational efficiency of Algorithm 1. In short, the proposed algorithm is feasible and effective in practical applications.

5. Analysis and discussion of C-PDDM

5.1. Analysis of convergence

To prove the convergence of Algorithm 1, the following lemma is proposed.

Lemma 1 (Nie et al., 2010). For any two non-zero vectors $V_1, V_2 \in \mathbb{R}^d$, the following inequality holds:

$$\|V_1\|_2 - \frac{\|V_1\|_2^2}{2\|V_2\|_2} \leq \|V_2\|_2 - \frac{\|V_2\|_2^2}{2\|V_2\|_2} \tag{24}$$

Then, we prove the convergence of the proposed algorithm through Theorem 5.**Theorem 5**. Algorithm 1 decreases the objective value of the optimization problem (17) in each iteration and converges to the optimal solution.

Proof. For expression simply, the updated results of optimization variables $\lambda_{k,c}$, W, and Y after τ-th iteration are denoted as $\lambda_{k,c}^\tau$, W^τ, and Y^τ, respectively. The internal loop iteration update in Step 8 of Algorithm 1 corresponds to the following optimization problem:

$$\begin{aligned}
&\Theta\left(\lambda_k, Y, W\right) \\
&= \min_{\lambda_{k,c}, Y, W} \sum_{c=0}^{C}\sum_{k=1}^{N} \lambda_{k,c}^2 \left\|\phi\left(x_{k,c}\right) - \mu_c\right\|_H^2 \\
&+ \beta \sum_{c=0}^{C}\sum_{k=1}^{N} \left(\lambda_{k,c}^2 ln \lambda_{k,c}^2 - \lambda_{k,c}^2\right) + \alpha tr(Y^T L Y) \\
&+ \sum_{c=0}^{C}\sum_{k=1}^{N} \lambda_{k,c}^2 \left\|\phi^T(x_k) W - y_k\right\|_H^2 + \rho tr\left(W^T U W\right)
\end{aligned} \tag{25}$$

According to the definition of matrix U, we have:

$$\begin{aligned}
&Z(\tau+1) + \rho \sum_{i=1}^{N} \frac{\left\|W(i,:)^{\tau+1}\right\|_2^2}{\left\|W(i,:)^{\tau}\right\|_2} \\
&\leq Z(\tau) + \rho \sum_{i=1}^{N} \frac{\left\|W(i,:)^{\tau}\right\|_2^2}{\left\|W(i,:)^{\tau}\right\|_2}
\end{aligned} \tag{26}$$

where

$$\begin{aligned}
Z(e) = &\sum_{c=0}^{C} \left(\lambda^{(c)^e}\right)^2 \left\|\phi\left(X^{(c)}\right) - \mu^{(c)}\right\|_H^2 \\
&+ \beta \sum_{c=0}^{C} \left(\left(\lambda^{(c)^e}\right)^2 ln\left(\lambda^{(c)^e}\right)^2 - \left(\lambda^{(c)^e}\right)^2\right) \\
&+ \alpha tr\left(\left(Y^e\right)^T L Y^e\right) \\
&+ \sum_{c=0}^{C} \left(\lambda^{(c)^e}\right)^2 \left\|\phi^T\left(X^{(c)}\right) W^e - Y^e\right\|_H^2
\end{aligned}$$

Based on Lemma 1, we can obtain the following inequality:

$$\begin{aligned}&\sum_{j=1}^{N}\left(\left\|(W)_{j,:}^{\tau+1}\right\|_2-\frac{\left\|(W)_{j,:}^{\tau+1}\right\|_2^2}{2\left\|(W)_{j,:}^{\tau}\right\|_2}\right)\\&\leq\sum_{j=1}^{N}\left(\left\|(W)_{j,:}^{\tau}\right\|_2-\frac{\left\|(W)_{j,:}^{\tau}\right\|_2^2}{2\left\|(W)_{j,:}^{\tau}\right\|_2}\right)\end{aligned}\tag{27}$$

Therefore, we can derive:

$$\begin{aligned}&\mathrm{Z}(\tau+1)+\rho\sum_{i=1}^{N}\left\|W_{j,:}^{\tau+1}\right\|_2\\&\leq\mathrm{Z}(\tau)+\rho\sum_{i=1}^{N}\left\|W_{j,:}^{\tau}\right\|_2\end{aligned}\tag{28}$$

Finally, Theorem 6 is proved.

According to the update rule in Algorithm 1 and Theorem 6, it is known that the optimization objective (17) is a decreasing function concerning the objective value. Therefore, it can be inferred that Algorithm 1 can effectively converge to the optimal solution.

5.2. Analysis of generalization

Rademacher complexity can effectively measure the ability of a function set to fit noise (Ghifary et al., 2017; Tao and Dan, 2021). Therefore, we will derive the generalization error bound of the proposed method through Rademacher complexity. Let $H:=\{h:\mathcal{X}\to\mathcal{Y}\}$ be a set of hypothesis functions in the RKHS $\mathcal{H}$ space, where $\boldsymbol{\mathcal{X}}$ is a compact set and $\boldsymbol{\mathcal{Y}}$ is a label space. Given a loss function $loss(\cdot,\cdot):\boldsymbol{\mathcal{Y}}\times\boldsymbol{\mathcal{Y}}\to\mathbb{R}_+$and a. neighborhood distribution $\boldsymbol{\mathcal{D}}$ on $\boldsymbol{\mathcal{X}}$, the expected loss of two hypothesis functions $h,\tilde{h}\in H$ is defined as:

$$\mathcal{L}_{\mathcal{D}}\left(h,\tilde{h}\right)=E_{x\sim\mathcal{D}}\left[loss\left(h(x),\tilde{h}(x)\right)\right]$$

The domain distribution difference between the source domain distribution $\mathbb{P}$ and the target domain distribution $\mathbb{Q}$ can be defined as:

$$disc(\mathbb{P},\mathbb{Q})=\sup_{h,\tilde{h}\in H}\left\{\mathcal{L}_{\mathbb{P}}\left(h,\tilde{h}\right)-\mathcal{L}_{\mathbb{Q}}\left(h,\tilde{h}\right)\right\}\tag{29}$$

Let $f_{\mathbb{P}}$ and $f_{\mathbb{Q}}$ be the true label functions for $\mathbb{P}$ and $\mathbb{Q}$, respectively, and let the corresponding optimized hypothesis functions be:

$$\begin{aligned}h_{\mathbb{P}}^{*}&:=\operatorname*{argmin}_{h\in H}\mathcal{L}_{\mathbb{P}}\left(h,f_{\mathbb{P}}\right)\\h_{\mathbb{Q}}^{*}&:=\operatorname*{argmin}_{h\in H}\mathcal{L}_{\mathbb{Q}}\left(h,f_{\mathbb{Q}}\right)\end{aligned}$$

Their corresponding expected loss is denoted as $\mathcal{L}_{\mathbb{P}}\left(h_{\mathbb{Q}}^{*},h_{\mathbb{P}}^{*}\right)$. Our C-PDDM method achieves the empirical loss target of $\mathcal{L}_{\mathbb{P}}\left(h_{\mathbb{Q}}^{*},h_{\mathbb{P}}^{*}\right)$ through the objective function $R(Y,W)$.

The following theorem gives the generalization error bound of the proposed method:

Theorem 6 (Generalization Error Bound) (Nie et al., 2010). Let $H:=\{f\in\mathcal{H}:\mathcal{X}\to\mathbb{R},\|f\|_{\mathcal{H}}\leq1\ and\ \|f\|_{\infty}\leq r\}$ is a function set of RKHS $\mathcal{H}$. $X_{\boldsymbol{\mathcal{X}}}^{\mathbb{P}}=\left(x_1^s,\ldots,x_{n_s}^s\right)\sim\mathbb{P}$ and $X_{\boldsymbol{\mathcal{X}}}^{\mathbb{Q}}=\left(x_1^t,\ldots,x_{n_t}^t\right)\sim\mathbb{Q}$ are datasets of the source domain and the target domain, respectively. q-Lipschitz function loss is $loss(\cdot,\cdot):\boldsymbol{\mathcal{Y}}\times\boldsymbol{\mathcal{Y}}\to[0,q]$. When $a,b\in\boldsymbol{\mathcal{Y}}\times\boldsymbol{\mathcal{Y}}$, $\left|loss(a)-loss(b)\right|=q|a-b|$. The generalization error bound for any hypothesis function $h\in\mathcal{H}$ with a probability of at least $1-\delta$ of having Rademacher complexity $\Re_{X_{\mathcal{X}}^{\mathbb{P}}}(H)$ on $X_{\boldsymbol{\mathcal{X}}}^{\mathbb{P}}$ is:

$$\begin{aligned}&\mathcal{L}_{\mathbb{Q}}\left(h,f_{\mathbb{Q}}\right)-\mathcal{L}_{\mathbb{Q}}\left(h_{\mathbb{Q}}^{*},f_{\mathbb{Q}}\right)\leq\mathcal{L}_{\mathbb{P}}\left(h,h_{\mathbb{P}}^{*}\right)+2q\Re_{X_{\mathcal{X}}^{\mathbb{P}}}(H)\\&+3q\sqrt{\frac{\log\frac{2}{\delta}}{2N}}+8q\sqrt{\Omega\left(\lambda_k,X_s,X_t\right)}+R(Y,W)\end{aligned},\tag{30}$$

where $\Re_{X_{\mathcal{X}}^{\mathbb{P}}}(H)$ is Rademacher complexity.

Theorem 6 shows that the possibilistic distribution distance measure $\Omega\left(\lambda_k,X_s,X_t\right)$ and the model alignment function $R(Y,\mathrm{W})$ can simultaneously control the generalization error bound of the proposed method. Therefore, the proposed method can effectively improve its generalization performance in domain adaptation by minimizing both the possibilistic distribution distance between domains and model bias. The experimental results on real-world datasets also confirm this conclusion.

5.3. Discussion of kernel selection

The literature (32) theoretically analyzed and pointed out that the Gaussian kernel cluster provides an effective RKHS embedding space for the consistency estimation of domain distribution distance measure. The detailed derivation process can be found in Sriperumbudur et al. (2010a,b). Therefore, all the kernel functions used in this paper are Gaussian kernel $k_\sigma=e^{-x_i-x_{j2}^2/2\sigma^2}$. In order to illustrate the impact of the Gaussian kernel bandwidth on the distribution of sample RKHS embedding, the following theorem is introduced:

Theorem 7 (Sriperumbudur et al., 2010a). The function set of Gaussian kernel.

$$\begin{aligned}K_s=\{&k_\sigma=e^{-\|x_i-x_j\|_2^2/2\sigma^2},x_i,x_j\in\mathbb{R}^d,\\&\sigma\in[\sigma_0,\infty),\sigma_0>0\}\end{aligned}\tag{31}$$

For any $\mathrm{k}_\sigma,\mathrm{k}_\theta\in K_s$ and $0<\theta<\sigma<\infty$, then $\zeta_{k\sigma}\left(X^s,X^t\right)\geq\zeta_{k\theta}\left(X^s,X^t\right)$.

According to Theorem 7, the larger the kernel bandwidth, the larger the RKHS embedding distance of the domain distribution, which slows down the convergence speed of the domain distribution distance measure $\Omega\left(\lambda_k,X_s,X_t\right)$ based on the soft clustering hypothesis of the MMD criterion. In order to further study the performance impact of Gaussian kernel bandwidth, the Gaussian kernel bandwidth is parameterized, that is, the generalized Gaussian kernel function is defined as:

$$\mathrm{k}_{\sigma/\theta}\left(x,X_i\right)=\exp\left(-\|x-X_i\|_2^2/2(\sigma/\theta)^2\right)\tag{32}$$

where θ is a tunable parameter, as will be shown in the experimental analysis below. When θ is too large, the samples within the domain are highly cohesive, leading to a certain degree of mixing between positive and negative classes, which is not conducive to effective classification of the model. Conversely, when θ is too small, it may slow down the convergence of the distribution distance measurement algorithm based on the possibilistic clustering hypothesis to some extent. Therefore, this paper limits $\theta \in [1, \theta_0]$, where θ_0 is a sufficiently large tunable parameter. The above analysis shows that the distribution distance measurement based on the possibilistic clustering hypothesis can not only constrain the divergence of the distributions between domains to be as consistent as possible, but also reduce the divergence of the sample distributions within each domain within a certain range of kernel bandwidths, thereby accelerating the convergence speed of the domain distribution divergence difference measurement and further improving the execution efficiency of the algorithm.

It is worth noting that kernel selection is an open problem in kernel learning methods. Recently, some studies have proposed the use of Multi-Kernel Learning (MKL) (Long et al., 2015) to overcome the kernel selection problem in single-kernel learning methods. Therefore, we can also use MKL to improve the performance of the proposed method. Specifically, the first step is to construct a new space that spans multiple kernel feature mappings, represented by $\{\phi_a\}_{a=1}^{\mho}$, which projects X into $\mho$ different spaces. Then, an orthogonal integration space can be built by connecting these $\mho$ spaces, and $\tilde{\phi}(x_i) = \left[\phi_1(x_i)^T, \phi_2(x_i)^T, \ldots, \phi_{\mho}(x_i)^T\right]^T \in \mathbb{R}^{\mho N}$ represents the mapping features in the final space, where $x_i \in X$. In addition, the kernel matrix in this final space can be written as $K_{new} = \left[\tilde{K}_1; \tilde{K}_2; \ldots; \tilde{K}_{\mho}\right]$, where $\widetilde{K_i}$ is the i-th kernel matrix from $\mho$ feature spaces. The kernel functions that can be used in practice include the Gaussian kernel function, inverse square distance kernel function $K_{ij} = 1/\left(1 + \sigma \|x_i - x_j\|^2\right)$, Laplacian kernel function $\mathbf{K}_{ij} = \exp\left(-\sqrt{\sigma}\|x_i - x_j\|\right)$, and inverse distance kernel function $K_{ij} = 1/\left(1 + \sqrt{\sigma}\|x_i - x_j\|\right)$, etc.

6. Experiments

6.1. Emotional databases and data preprocessing

In order to make a fair comparison with stat-of-the-art (SOTA) methods, a large number of experiments were conducted for effective validation on two well-known open datasets [i.e., SEED (Zheng and Lu, 2015) and SEED-IV (Zheng et al., 2019)]. The SEED dataset has a total of 15 subjects participating in the experiment to collect data, each subject needs to have three sessions at different times, each session contains 15 trials, with a total of 3 emotional stimuli (negative, neutral, and positive). In the SEED-IV dataset, there are also 15 subjects participating in the experiment to collect data, each subject needs to have three sessions at different times, each session contains 24 trials, with a total of 4 emotional stimuli (happy, sad, fearful, and peaceful).

The EEG signals of the two datasets (i.e., SEED and SEED-IV) are collected simultaneously from the 62-channel ESI Neuroscan system. In the EEG signal preprocessing, the down-sampled data sampling rate is reduced to 200 Hz, then the environmental noise data is manually removed, and the data is filtered through a 0.3 Hz–50 Hz band-pass filter. In each trial, the data is divided into multiple segments with a length of 1 s. Based on the predefined 5 frequency band-passes [Delta (1–3 Hz), Theta (4–7 Hz), Alpha (8–13 Hz), Beta (14–30 Hz), and Gamma (31–50 Hz)], the corresponding differential entropy (DE) is extracted to represent the logarithmic power spectrum in the specified frequency band-pass, and a total of 310 features (5 frequency bands and 62 channels) are obtained in each EEG segment. Then, all features are smoothed by the Linear Dynamic System (LDS) method, which can utilize the time dependency of emotion transitions and filter out the noise EEG components unrelated to emotions (Shi and Lu, 2010).

6.1.1. Settings

The settings of the hyper-parameter for the C-PDDM method are also crucial before analyzing the experimental evaluation results. For all methods, in both the source and target domains, a Gaussian kernel $K(x, x_i) = \exp(-\|x - x_i\|^2 / 2\sigma^2)$ is used, where σ can be obtained by minimizing MMD to obtain a benchmark test. Based on experience, we first select σ as the square root of the average norm of the binary training data, and $\sigma\sqrt{C}$ (where C is the number of classes) for multi-class classification. The underlying geometric structure depends on k neighbors to compute the Laplacian matrix. In the experiment of this paper, it can be observed that the performance slightly varies when k is not large. Therefore, to construct the nearest neighbor graph in C-PDDM, this paper conducts a grid search for the optimal number of nearest k neighbors in $\{3,5,10,15,17\}$, and provides the best recognition accuracy results from the optimal parameter configuration.

Before presenting the detailed evaluation, it is necessary to explain how the hyper-parameters of C-PDDM are tuned. Based on experience, the parameter β is used to balance the fuzzy entropy and domain probability distribution alignment in the objective function (16). Both parameters α and ρ are adjustable parameters, and they are used to balance the importance of structure description and feature selection. Therefore, these two parameters have a significant impact on the final performance of the method.

Considering that parameter uncertainty is still an open problem in the field of machine learning, we determine these parameters based on previous work experience. Therefore, we evaluate all methods on the dataset by empirically searching the parameter space to obtain the optimal parameter settings and give the best results for each method. Except for special cases, all parameters of all relevant methods are tuned to obtain the optimal results.

As unsupervised domain adaptation does not have target labels to guide standard cross-validation, we perform leave-one-subject-out on the two datasets: SEED and SEED-IV (the details of this protocol are shown in Section 6.2). We obtain the optimal parameter values on $\{10^{-6}, 10^{-5}, \ldots, 10^{5}, 10^{6}\}$ by obtaining the highest average accuracy on the two datasets using the above method. This strategy often constructs a good C-PDDM model for unsupervised domain adaptation, and a similar strategy is adopted to find the optimal parameter values for other domain adaptation methods. In the following sub-sections, a set of experiments is set up to test the sensitivity of the proposed method C-PDDM to parameter selection (i.e., Section 6.4.1), in order to verify that C-PDDM can achieve stable performance within a wide range of parameter values. In addition, the hyper-parameters of other methods are selected according to the original literature.

TABLE 1 The mean accuracies (%) and standard deviations (%) of emotion recognition on the SEED database using cross-subject cross-session leave-one-subject-out cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF (Breiman, 2001)	69.60 ± 7.64	KNN (Coomans and Massart, 1982)	60.66 ± 7.93
SVM* (Suykens and Vandewalle, 1999)	62.24 ± 5.48	Adaboost (Zhu et al., 2006)	71.87 ± 5.70
TCA* (Pan et al., 2011)	65.31 ± 6.04	CORAL (Sun et al., 2016)	69.22 ± 4.11
SA (Li Y. et al., 2020)	61.41 ± 9.75	GFK* (Gong et al., 2012)	67.36 ± 6.52
DICE* (Liang et al., 2018)	73.56 ± 4.23	C-PDDM	**73.82 ± 6.12**
Deep learning methods			
DCORAL* (Sun et al., 2016)	80.87 ± 6.04	DAN* (Long et al., 2015)	82.51 ± 3.71
DDC (Tzeng et al., 2014)	82.17 ± 4.96	DANN* (Ganin et al., 2016)	84.79 ± 6.44
PR-PL (Zhou et al., 2022)	85.56 ± 4.78	C-PDDM+ResNet101	**86.49 ± 5.20**

Here, the model results reproduced by us are indicated by "*". The bold values are the best performance in tables.

6.2. Experiment protocols

In order to fully verify the robustness and stability of the proposed method, we adopt four different validation protocols (leave-one-subject-out) (Zhang et al., 2021) to compare the proposed method with the SOTA methods.

1) **Cross-subject cross-session leave-one-subject-out cross-validation.** To fully estimate the robustness of the model on unknown subjects and trials, this paper uses a strict leave-one-out method cross-subject cross-session to evaluate the model. All session data of one subject is used as the target domain, and all sessions of the remaining subjects are used as the source domain. We repeat the training and validation until all sessions of each subject have been used as the target domain once. Due to the differences between subjects and sessions, this evaluation protocol poses a significant challenge to the effectiveness of models in emotion recognition tasks based on EEG.
2) **Cross-subject single-session leave-one-subject-out cross-validation.** This is the most widely used validation scheme in emotion recognition tasks based on EEG (Luo et al., 2018; Li J. et al., 2020). One session data of a subject is treated as the target domain, while the remaining subjects are treated as the source domain. We repeat the training and validation process until each subject serves as the target once. As with other studies, we only consider the first session in this type of cross-validation.
3) **Within-subject cross-session leave-one-session-out cross-validation.** Similar to existing methods, a time series cross-validation method is employed here, where past data is used to predict current or future data. For a subject, the first two sessions are treated as the source domain, and the latter session is treated as the target domain. The average accuracy and standard deviation across subjects are calculated as the final results.
4) **Within-subject single-session cross-validation.** Following the validation protocols proposed in existing studies (Zheng and Lu, 2015; Zheng et al., 2019), for each session of a subject, we take the first 9 (SEED) or 16 (SEED-IV) trials as the source domain and the remaining 6 (SEED) or 8 (SEED-IV) trials as the target domain. The results are reported as the average performance of all participants. In the performance comparison of the following four different validation protocols, we use "*" to indicate the replicated model results.

6.3. Results analysis on SEED and SEED-IV

6.3.1. Cross-subject cross-session

For verifying the efficiency and stability of the model under cross-subject and cross-session conditions, we used cross-subject cross-session leave-one-subject-out cross-validation on the SEED and SEED-IV databases to validate the proposed C-PDDM. As shown in Tables 1, 2, the results show that our proposed model achieved the highest accuracy of emotion recognition. The C-PDDM method, with or without using deep features, achieved emotion recognition performances of 73.82 ± 6.12 and 86.49 ± 5.20 for the three-class classification task on SEED, and 67.83 ± 8.06 and 72.88 ± 6.02 for the four-class classification task on SEED-IV. Compared with existing research, the proposed C-PDDM has a slightly lower accuracy on SEED-IV than PR-PL, but PR-PL uses adversarial learning, which has a higher computational cost. In addition, the proposed C-PDDM method has the best recognition performance in the other three cases. These results indicate that the proposed C-PDDM has a higher recognition accuracy and better generalization ability, and is more effective in emotion recognition.

6.3.2. Cross-subject single-session

Table 3 summarizes the model results of the recognition task under cross-subject single-session leave-one-subject-out and compares them with the performance of the latest methods in the literature. All results are presented in the form of mean ± standard deviation. The results show that our proposed model (C-PDDM) achieves the best performance (74.92%) with a standard deviation of 8.16 when compared with traditional machine learning methods. The recognition performance of C-PDDM is better than the DICE method, indicating that the C-PDDM method is superior to the DICE method in dealing with noisy situations. When compared with the latest deep learning

TABLE 2 The mean accuracies (%) and standard deviations (%) of emotion recognition on SEED-IV database using cross-subject cross-session leave-one-subject-out cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF	50.98 ± 9.20	KNN	40.83 ± 7.28
SVM	51.78 ± 12.85	Adaboost	53.44 ± 9.12
TCA	56.56 ± 13.77	CORAL	49.44 ± 9.09
SA	64.44 ± 9.46	GFK	45.89 ± 8.27
KPCA (Suykens and Vandewalle, 1999)	51.76 ± 12.89	DNN (Suykens and Vandewalle, 1999)	49.35 ± 9.74
DICE	66.75 ± 7.25	C-PDDM	**67.83 ± 8.06**
Deep learning methods			
DGCNN (Song et al., 2018)	52.82 ± 9.23	DAN	58.87 ± 8.13
RGNN (Zhong et al., 2020)	73.84 ± 8.02	BiHDM (Li Y. et al., 2020)	69.03 ± 8.66
BiDANN (Li et al., 2018c)	65.59 ± 10.39	DANN	54.63 ± 8.03
PR-PL	**74.92 ± 7.92**	C-PDDM+ResNet101	72.88 ± 6.02

The bold values are the best performance in tables.

TABLE 3 The mean accuracies (%) and standard deviations (%) of emotion recognition on the SEED database using cross-subject single-session leave-one-subject-out cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
TKL (Li et al., 2018c)	63.54 ± 15.47	T-SVM* (Li et al., 2018c)	68.57 ± 9.54
TCA	63.64 ± 14.88	TPT* (Suykens and Vandewalle, 1999)	73.86 ± 11.05
KPCA	61.28 ± 14.62	GFK	71.31 ± 14.09
SA*	66.00 ± 10.89	DICA (Ma et al., 2019)	69.40 ± 07.80
DNN	61.01 ± 12.38	SVM	58.18 ± 13.85
DICE	74.22 ± 7.33	C-PDDM	**74.92 ± 8.16**
Deep learning methods			
DGCNN	79.95 ± 9.02	DAN	83.81 ± 8.56
RGNN	85.30 ± 6.72	BiHDM	85.40 ± 7.53
WGAN-GP (Luo et al., 2018)	87.10 ± 7.10	MMD (Li J. et al., 2020)	80.88 ± 10.10
ATDD-DANN (Du et al., 2020)	90.92 ± 1.05	JDA-Net (Li J. et al., 2020)	88.28 ± 11.44
R2G-STNN (Li et al., 2019)	84.16 ± 7.63	SimNet* (Pinheiro, 2018)	81.58 ± 5.11
BiDANN	83.28 ± 9.60	DResNet (Ma et al., 2019)	85.30 ± 8.00
ADA (Li J. et al., 2020)	84.47 ± 10.65	DANN	81.65 ± 9.92
PR-PL	**93.06 ± 5.12**	C-PDDM+ResNet101	92.19 ± 4.70

Here, the model results reproduced by us are indicated by "*". The bold values are the best performance in tables.

methods, especially with deep transfer learning networks based on DANN (Li J. et al., 2020) [such as ATDD-DANN (Du et al., 2020), R2GSTNN(Li et al., 2019), BiHDM (Li Y. et al., 2020), BiDANN (Li et al., 2018c), WGAN-GP (Luo et al., 2018)], the proposed C-PDDM method effectively addresses individual differences and noisy label issues in aBCI applications. The recognition performance of PR-PL is slightly better than the C-PDDM, which may be because the PR-PL method uses adversarial loss for model learning, resulting in higher computational costs. Overall, the C-PDDM method has a competitive result, indicating that the C-PDDM method has better generalization performance in cross-subject within the same session.

6.3.3. Within-subject cross-session

By calculating the mean and standard deviation of the experimental results for each subject, the cross-session cross-validation results for each subject on the different datasets SEED and SEED-IV are shown in Tables 4, 5, respectively. For these two datasets, our proposed C-PDDM method, which compared with the existing traditional machine learning methods, has results close to or better than the DICE method on both SEED and SEED-IV. This may be because each subject is less likely to generate noisy data in different sessions, which does not highlight the advantages of C-PDDM. In addition, for the SEED-IV dataset (four-class emotion recognition), regardless of traditional machine learning or the latest deep learning methods, the performance of the C-PDDM method is the best when the number of categories increases. This indicates that the proposed method is more accurate and has stronger scalability in more nuanced emotion recognition tasks.

TABLE 4 The mean accuracies (%) and standard deviations (%) of emotion recognition on the SEED database using within-subject cross-session cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF	76.42 ± 11.15	KNN*	72.96 ± 12.10
TCA*	77.63 ± 11.49	CORAL	84.18 ± 9.81
SA*	67.79 ± 7.43	GFK*	79.28 ± 7.44
DICE	**81.58 ± 7.55**	C-PDDM	**81.58 ± 9.30**
Deep learning methods			
DAN	89.16 ± 7.90	SimNet	86.88 ± 7.83
DDC	91.14 ± 5.61	ADA	89.13 ± 7.13
DANN	89.45 ± 6.74	MMD	84.38 ± 12.05
JDA-Net	91.17 ± 8.11	DCORAL (Sun et al., 2016)	88.67 ± 6.25
PR-PL	**93.18 ± 6.55**	C-PDDM+ResNet101	92.56 ± 5.29

The bold values are the best performance in tables.

TABLE 5 The mean accuracies (%) and standard deviations (%) of emotion recognition on SEED-IV database using within-subject cross-session cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
RF	60.27 ± 16.36	KNN	54.18 ± 16.28
TCA*	59.49 ± 12.07	CORAL*	66.88 ± 14.67
SA*	56.94 ± 11.45	GFK*	60.66 ± 10.00
DICE	69.68 ± 12.52	C-PDDM	**70.48 ± 9.08**
Deep learning methods			
DCORAL (Chen et al., 2021)	65.10 ± 13.20	DAN	60.20 ± 10.20
DDC (Chen et al., 2021)	68.80 ± 16.60	MEERNet (Chen et al., 2021)	72.10 ± 14.10
PR-PL	74.62 ± 14.15	C-PDDM+ResNet101	**76.29 ± 11.36**

The bold values are the best performance in tables.

TABLE 6 The mean accuracies (%) and standard deviations (%) of emotion recognition on the SEED database using within-subject single-session cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
SVM*	77.80 ± 12.61	GRSLR (Li et al., 2018a)	**87.39 ± 8.64**
RF	78.46 ± 11.77	GSCCA (Zheng, 2017)	82.96 ± 9.95
CCA	77.63 ± 13.21	DBN (Zheng et al., 2015)	86.08 ± 8.34
DICE	86.28 ± 9.22	C-PDDM	86.74 ± 7.59
Deep learning methods			
DGCNN	90.40 ± 8.49	RGNN	94.24 ± 5.95
ATDD-DANN	91.08 ± 6.43	BiHDM	93.12 ± 6.06
R2G-STNN	93.38 ± 5.96	SimNet*	90.13 ± 10.84
BiDANN	92.38 ± 7.04	STRNN (Zhang et al., 2019a)	89.50 ± 7.63
GCNN (Breiman, 2001)	87.40 ± 9.20	DANN	91.36 ± 8.30
PR-PL	94.84 ± 9.16	C-PDDM+ResNet101	**96.38 ± 6.88**

The bold values are the best performance in tables.

6.3.4. Within-subject single-session

The previous evaluation strategy only considered the first two sessions of the SEED dataset as the source domain for the experiment. The evaluation results of emotion recognition for each subject within each session are presented in Table 6. When compared with traditional machine learning methods, the C-PDDM method has comparable

TABLE 7 The mean accuracies (%) and standard deviations (%) of emotion recognition on SEED-IV database using within-subject single-session cross-validation.

Methods	Pacc	Methods	Pacc
Traditional machine learning methods			
SVM	56.61 ± 20.05	GRSLR	69.32 ± 19.57
RF	50.97 ± 16.22	GSCCA	69.08 ± 16.66
CCA	54.47 ± 18.48	DBN	66.77 ± 07.38
DICE	71.67 ± 11.29	C-PDDM	**71.85 ± 9.18**
Deep learning methods			
DGCNN	69.88 ± 16.29	RGNN	79.37 ± 10.54
GCNN	68.34 ± 15.42	BiHDM	74.35 ± 14.09
A-LSTM (Breiman, 2001)	69.50 ± 15.45	SimNet*	71.38 ± 13.12
BiDANN	70.29 ± 12.63	DANN	63.07 ± 12.66
PR-PL	83.33 ± 10.61	C-PDDM+ResNet101	**83.94 ± 11.39**

The bold values are the best performance in tables.

performance, and it still outperforms the performance of the DICE method. When compared with the latest deep learning methods, the C-PDDM method achieves the highest recognition performance, reaching 96.38%, which is even higher than the PR-PL method. This comparison demonstrates the high efficiency and reliability of the proposed C-PDDM method in various emotion recognition applications.

For the SEED-IV dataset, we calculated the performance of all three sessions (emotional categories: happiness, sadness, fear, and neutral). Our proposed model outperforms the existing latest classical research methods and achieves the highest accuracy of 71.85 and 83.94% in Table 7. This comparison shows that the more emotional categories there are, the more prominent the generalization of the proposed C-PDDM method in applications.

6.4. Discussion

For comprehensively study the performance of the model, we evaluated the effects of different settings in C-PDDM. Please note that all the results presented in this section are based on the SEED dataset, using the cross-subject single-session cross-validation evaluation protocol.

6.4.1. Ablation study

We conducted ablation studies to systematically explore the effectiveness of different components in the proposed C-PDDM model and their respective contributions to the overall performance of the model. As shown in Table 8, when 5 labeled samples existed at each category in the target domain, the recognition accuracy (93.83% ± 5.17) is very close to the recognition accuracy of C-PDDM (unsupervised learning) (92.19% ± 4.70). This decrease indicates the impact of individual differences on model performance and highlights the huge potential of transfer learning in aBCI applications. Moreover, the results show that simultaneously preserving the local structure of data in both the source and target domains helps improve model performance; otherwise, the recognition accuracy decreases significantly (90.60% ± 5.29 and 91.37% ± 5.82, respectively). When $\|W\|_{2,1}$ is changed to $\|W\|_2$, the model's recognition accuracy drops to 91.84% ± 6.33. This result reflects the sample selection and denoising effects achieved when using $l_{2,1}$ constraint.

TABLE 8 The ablation study of our proposed model.

Ablation study about training strategy	Pacc
target prior information (5 labeled samples per category)	93.83 ± 5.17
only preserving the local structures on the source	90.60 ± 5.29
only preserving the local structures on the target	91.37 ± 5.82
imposing l_2-norm on W	91.84 ± 6.33
fixed pseudo-labeling	89.95 ± 5.61
dynamic pseudo-labeling	92.19 ± 4.75
multiple kernel leaning	93.68 ± 6.04
Hyper-parameter controlling strategy	
$\alpha = 0$ (ignoring the local structures)	90.27 ± 5.51
fixed $\alpha = 1$ for local preserving regularization	91.93 ± 5.44
fixed $\beta = 100$ for fuzzy entropy regularization	92.17 ± 6.30
fixed ρ for W regularization	92.16 ± 5.38
$\delta = 0$	88.47 ± 6.00
$\delta = 0.3$	88.91 ± 3.49
$\delta = 0.5$	92.19 ± 4.70
$\delta = 0.85$	91.83 ± 2.80
$\delta = 1$	89.85 ± 5.66
$\beta = 0$ (ignoring the fuzzy entropy regularization)	90.56 ± 6.59
The proposed model	
C-PDDM+ResNet101	92.19 ± 4.70

For the pseudo-labeling method, when the pseudo-labeling method changes from fixed to linear dynamic, the corresponding accuracy increases from 89.95 to 92.19%. When adopting multi-kernel learning, the accuracy further improves to 93.68%. The results indicate that multi-kernel learning helps rationalize the importance of each kernel in different scenarios and enhances the generalization of the model.

Next, we analyze the impact of different hyper-parameters on the overall performance of the model. According to the experimental results, it can be seen that the recognition accuracy with α, β, ρ are dynamically learned better than fixed values. When ignoring the local

structural information and fuzzy entropy information in the domain, the performance drops by about 2% (i.e., $\alpha = 0$, $\alpha = 1$, $\beta = 0$, and $\beta = 100$). In addition, from the results, it can be inferred that the performance is optimal when the value of δ is around 0.5, indicating that the means of different categories in the source domain and target domain are equally important.

6.4.2. Effect of noisy labels

In order to further verify the robustness of the model in the noisy label learning process, we randomly add noise to the source labels at different ratios and test the performance of the corresponding model on unknown target data. Specifically, we replace the corresponding proportion of real labels in Y^S with randomly generated labels to train the model by semi-supervised learning and then test the performance of the trained model in the target domain. It should be noted that only noise data is added in the source domain, and the target domain needs to be used for model evaluation. In the implementation, the noise ratios are adjusted to 5, 15, 25, and 30% of the sample number of the source domain, respectively. The results in Figure 2 show that the accuracy of the proposed C-PDDM decreases at the slowest rate as the number of noise increases. It indicates that C-PDDM is a reliable model with a high tolerance to noisy data. In future work, we can combine recently proposed new methods, such as Xiao et al. (2020) and (Jin et al. (2021), to further eliminate more common noise in EEG signals and improve the stability of the model in cross-corpus applications.

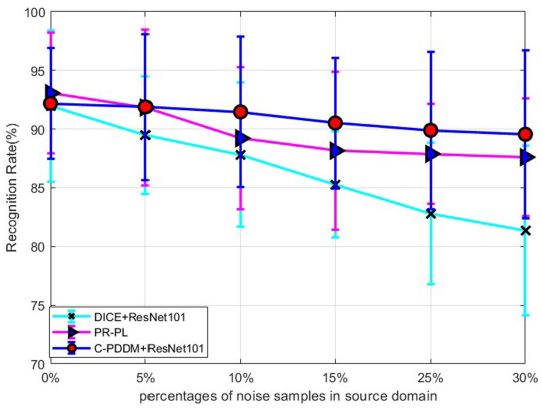


FIGURE 2
Robustness on source domain with different noise levels.

6.4.3. Confusion matrices

In order to qualitatively study the performance of the model in each emotion category, we analyze the confusion matrix through visualization and compare the results with the latest models (i.e., BiDANN, BiHDM, RGNN, PR-PL, DICE ResNet101). As shown in Figure 3, all models are good at distinguishing positive emotions from other emotions (with recognition rates above 90%), but relatively not good at distinguishing negative emotions and neutral emotions. For example, the emotion recognition rate in BiDANN (Li et al., 2018c) is even lower than 80% (76.72%). In addition, the PR-PL method achieves the best performance, possibly due to its adoption of adversarial networks, but at the cost of increased computational expenses. Compared with other existing methods (Figures 3A–C,E), our proposed model can improve the model's recognition ability, especially in distinguishing neutral and negative emotions, and its overall performance is better than the DICE method (as shown in Figures 3E,F).

6.4.4. Convergence

The proposed C-PDDM adopts an iterative optimization strategy and uses experiments to prove its convergence. The experiment is completed on the MATLAB platform, and the device configuration used is as follows: 64 GB memory, 2.5 GHz CPU, and 8-core Intel i7-11850H processor. Figure 4 shows the convergence process of C-PDDM at different iteration times. The results are shown in Figure 4. We can observe clearly that the proposed algorithm can achieve the minimum convergence at about 30 iterations. In the algorithm, the objective function of optimizing the sub-problem at

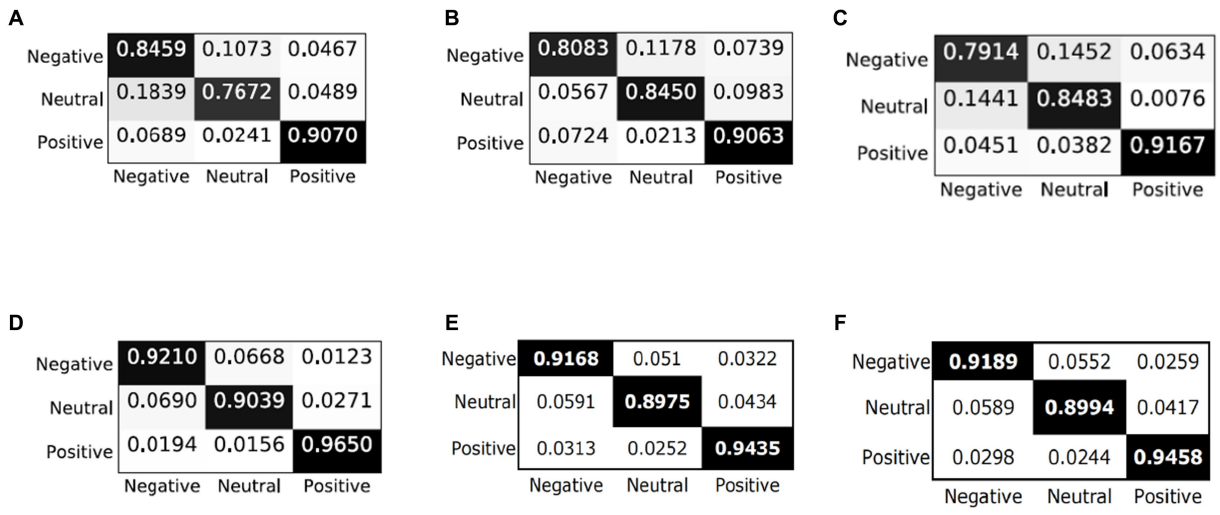


FIGURE 3
Confusion matrices of different models: **(A)** BiDANN; **(B)** BiHDM; **(C)** RGNN; **(D)** PR-PL; **(E)** DICE+ResNet101; and **(F)** C-PDDM+ResNet101.

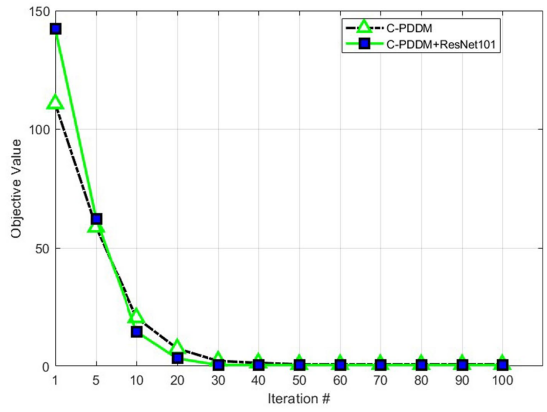


FIGURE 4
Convergence.

each time is a decreasing function, which proves that the C-PDDM method has good convergence.

7. Conclusion

This paper proposes a novel transfer learning framework based on a Clustering-based Probability Distribution Distance Metric (C-PDDM) hypothesis, which uses a probability distribution distance metric criterion and fuzzy entropy technology for EEG data distribution alignment, and introduces the Laplace matrix to preserve the local structural information of source and target domain data. We evaluate the proposed C-PDDM model on two famous emotion databases (SEED and SEED-IV) and compare it with existing state-of-the-art methods under four cross-validation protocols (cross-subject single-session, single-subject single-session, single-subject cross-session, and cross-subject cross-session). Our extensive experimental results show that C-PDDM achieves the best results in most of the four cross-validation protocols, demonstrating the advantages of C-PDDM in dealing with individual differences and noisy label issues in aBCI systems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

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

Funding

This work was supported by the Ningbo Natural Science Foundation (project no. 2022J180).

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

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. (2013). "Unsupervised domain adaptation by domain invariant projection". *In Proc. the 2013 IEEE International Conference on Computer Vision*, 769–776.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn.* 79, 151–175. doi: 10.1007/s10994-009-5152-4

Breiman, (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bruzzone, L., and Marconcini, M. (2010). Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 770–787. doi: 10.1109/TPAMI.2009.57

Carlucci, F. M, Porzi, L, Caputo, B, Ricci, E, et al. (2017). Autodial: Automatic domain alignment layers. In: Proceeding of 2017 IEEE international conference on computer vision (ICCV), Venice, pp: 5077–5085.

Chen, H., Li, Z., Jin, M., and Li, J. (2021). "Meernet: multi-source EEG-based emotion recognition network for generalization across subjects and sessions" in *43rd annual international conference of the IEEE engineering in Medicine & Biology Society (EMBC)*, vol. *2021* (IEEE), 6094–6097.

Chen, Z L, Zhang, J Y, Liang, X D, and Lin, L. Blending-target domain adaptation by adversarial meta-adaptation networks. In: Proceeding of 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), June 15-20, Long Beach (2019).

Chu, W.-S., Torre, F. D. L., and Cohn, J. F. (2013). "Selective transfer machine for personalized facial action unit detection" in *Proceeding of 2013 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (Portland, OR), 3515–3522.

Coomans, D., and Massart, L. D. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition: part 1. K-nearest neighbour classification by using alternative voting rules. *Anal. Chim. Acta* 136, 15–27. doi: 10.1016/S0003-2670(01)95359-0

Dan, Y., Tao, J., Fu, J., and Zhou, D. (2021). Possibilistic clustering-promoting semi-supervised learning for EEG-based emotion recognition. *Front. Neurosci.* 15:690044. doi: 10.3389/fnins.2021.690044

Dan, Y., Tao, J., and Zhou, D. (2022). Multi-model adaptation learning with possibilistic clustering assumption for EEG-based emotion recognition. *Front. Neurosci.*:16. doi: 10.3389/fnins.2022(16):855421

Ding, Z. M., Li, S., Shao, M., and Fu, Y. (2018). "Graph adaptive knowledge transfer for unsupervised domain adaptation" in *European Proceeding of conference on computer vision* (Munich), 36–52.

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science* 298, 1191–1194. doi: 10.1126/science.1076358

Du, X., Ma, C., Zhang, G., Li, J., Lai, Y. K., Zhao, G., et al. (2020). An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.*:1. doi: 10.1109/TAFFC.2020.3013711

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.48550/arXiv.1505.07818

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2017). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Trans. Patt. Anal. Mach. Intell.* 99:1. doi: 10.48550/arXiv.1510.04373

Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. *IEEE Conf. Comput. Vis. Patt. Recogn.* 2012, 2066–2073. doi: 10.1109/CVPR.2012.6247911

Gretton, A., Borgwardt, K. M., Rasch, M., Scholkopf, B., and Smola, A. J. (2007). "A kernel method for the two-sample-problem" in *Proceeding of the 21st annual conference on neural information processing systems, December 3-6* (Vancouver, BC).

Gretton, A, Harchaoui, Z, Fukumizu, K J, Harchaoui, Z, and Sriperumbudur, BK (2010). A fast, consistent kernel two-sample test. In: Proceedings of the 22nd international conference on neural information processing systems. 673–681. (Vancouver, BC, Canada).

Jayaram, V., Alamgir, M., Altun, Y., Scholkopf, B., and Grosse-Wentrup, M. (2016). Transfer learning in brain-computer interfaces abstract. The performance of brain-computer interfaces (BCIs) improves with the amount of avail. *IEEE Comput. Intell. Mag.* 11, 20–31. doi: 10.1109/MCI.2015.2501545

Jenke, R., Peer, A., and Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Trans. Affect. Comput.* 5, 327–339. doi: 10.1109/TAFFC.2014.2339834

Jin, J., Xiao, R., Daly, I., Miao, Y., Wang, X., and Cichocki, A. (2021). Internal feature selection method of CSP based on L1-norm and dempster–Shafer theory. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4814–4825. doi: 10.1109/TNNLS.2020.3015505

Kang, G. L., Jiang, L., Wei, Y., Yang, Y., and Hauptmann, A. (2022). Contrastive adaptation network for single- and multi-source domain adaptation. *Inst. Elect. Electron. Eng. Trans. Patt. Anal. Mach. Intell.* 44, 1793–1804. doi: 10.1109/TPAMI.2020.3029948

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Comput. Math. Methods Med.* 2013:573734. doi: 10.1155/2013/573734

Krishnapuram, R., and Keller, J.-M. (1993). A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst.* 1, 98–110. doi: 10.1109/91.227387

Lan, Z., Sourina, O., Wang, L., Scherer, R., and Muller-Putz, G. R. (2019). Domain adaptation techniques for EEG-based emotion recognition: a comparative study on two public datasets. *IEEE Trans. Cogn. Dev. Syst.* 11, 85–94. doi: 10.1109/TCDS.2018.2826840

Lee, S M, Kim, D W, Kim, N, and Jeong, SG. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In: Proceeding of 2019 IEEE/CVF international conference on computer vision (ICCV), October 27-November 2, Seoul (2019). pp: 90–100.

Li, J., Qiu, S., du, C., Wang, Y., and He, H. (2020). Domain adaptation for EEG emotion recognition based on latent representation similarity. *IEEE Trans. Cogn. Dev. Syst.* 12, 344–353. doi: 10.1109/TCDS.2019.2949306

Li, H., Jin, Y. M., Zheng, W. L., and Lu, B. L. (2018d). "Cross-subject emotion recognition using deep adaptation networks" in *Neural information processing*. eds. L. Cheng, A. C. S. Leung and S. Ozawa (Cham: Springer International Publishing), 403–413.

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring EEG features in cross-subject emotion recognition. *Front. Neurosci.* 12:162. doi: 10.3389/fnins.2018.00162

Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L., Cui, Z., et al. (2020). A novel bi-hemispheric discrepancy model for EEG emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* 13, 354–367. doi: 10.1109/TCDS.2020.2999337

Li, Y., Zheng, W., Cui, Z., Zhang, T., and Zong, Y. A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition. The 27th international joint conference on artificial intelligence (IJCAI) (2018b).

Li, Y., Zheng, W., Cui, Z., Zong, Y., and Ge, S. (2018a). EEG emotion recognition based on graph regularized sparse linear regression. *Neural. Process. Lett.* 49, 555–571. doi: 10.1007/s11063-018-9829-1

Li, Y., Zheng, W., Wang, L., Zong, Y., and Cui, Z. (2019). From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2019.2922912

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018c). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 12, 494–504. doi: 10.1109/TAFFC.2018.2885474

Liang, J., He, R., Sun, Z. N., and Tan, T. (2018). Aggregating randomized clustering-promoting invariant projections for domain adaptation. *Inst. Electr. Electron. Eng. Trans. Patt. Anal. Mach. Intell.* 41, 1027–1042. doi: 10.1109/TPAMI.2018.2832198

Long, M., Cao, Y., Wang, J., and Jordan, M., Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd international conference on international conference on machine learning, Lille, 97–105 (2015).

Long, M S, Wang, J M, Ding, G G, Sun, J, and Yu, PS. Transfer feature learning with joint distribution adaptation. In: Proceedings of the 2013 IEEE international conference on computer vision. IEEE, (2013).

Long, M. S., Wang, J. M., and Jordan, M. I. (2016). "Unsupervised domain adaptation with residual transfer networks" in *Proceeding of the 30th Annual conference on neural information processing systems, December 5-10* (Barcelona), 136–144.

Luo, L. K., Chen, L. M., Hu, S. Q., Lu, Y., and Wang, X. (2020). Discriminative and geometry aware unsupervised domain adaptation. *IEEE Trans. Cybern.* 50, 3914–3927. doi: 10.1109/TCYB.2019.2962000

Luo, Y., Zhang, S. Y., Zheng, W. L., and Lu, BL. Wgan domain adaptation for EEG-based emotion recognition, In: International Conference on Neural Information Processing (2018).

Ma, B.-Q., Li, H., Zheng, W.-L., and Lu, B.-L. (2019). "Reducing the subject variability of eeg signals with adversarial domain generalization" in *Neural information processing*. eds. T. Gedeon, K. W. Wong and M. Lee (Cham: Springer International Publishing), 30–42.

Mühl, C., Allison, B., Nijholt, A., and Chanel, G. (2014). A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain Comput. Interfaces* 1, 66–84. doi: 10.1080/2326263X.2014.912881

Musha, T., Terasaki, Y., Haque, H. A., and Ivamitsky, G. A. (1997). Feature extraction from EEGs associated with emotions. *Artif. Life Robot.* 1, 15–19. doi: 10.1007/BF02471106

Nie, F P, Huang, H, Cai, X, and Huang, H. Efficient and robust feature selection via joint -norms minimization. In: Proceedings of the 23rd international conference on neural information processing systems. Curran Associates Inc (2010): 1813–1821.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* 22, 199–210. doi: 10.1109/TNN.2010.2091281

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

Pandey, P., and Seeja, K. "Emotional state recognition with EEG signals using subject independent approach" Lecture notes on data engineering and communications technologies, data science and big data analytics, (Springer) (2019) 117–124. doi: 10.1.7/978-981-10-7641-1_10

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.* 32, 53–69. doi: 10.1109/MSP.2014.2347059

Pinheiro, P. O. (2018). Unsupervised domain adaptation with similarity learning. *IEEE/CVF Conf. Comput. Vis. Patt. Recogn.* 2018, 8004–8013. doi: 10.48550/arXiv.1711.08995

Shi, L.-C., and Lu, B.-L. (2010). Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning. *Annu. Int. Conf. IEEE Eng. Med. Biol.* 2010, 6587–6590. doi: 10.1109/IEMBS.2010.5627125

Song, T., Zheng, W., Song, P., and Cui, Z. (2018). EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* 11:1. doi: 10.1109/BIBM.2018.8621147

Sriperumbudur, B K, Fukumizu, K, Gretton, A, GRG, Lanckriet, and Scholkopf, B. Kernel choice and classifiability for RKHS embeddings of probability distributions. In: Proceeding of the 23rd annual conference on neural information processing systems (NIPS 2009). Red Hook, NY: MIT Press, 2010:1750–1758 (2010a).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., GRG, L., and Scholkopf, B. (2010b). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* 11, 1517–1561. doi: 10.1007/s10846-009-9337-7

Sun, B., Feng, J., and Saenko, K., Return of frustratingly easy domain adaptation. In: Proceedings of the thirtieth AAAI conference on artificial intelligence, ser. AAAI'16. AAAI Press, (2016), p. 2058–2065.

Sun, Y., Gao, Y., Zhao, Y., Liu, Z., Wang, J., Kuang, J., et al. (2022). Neural network-based tracking control of uncertain robotic systems: predefined-time nonsingular terminal sliding-mode approach. *IEEE Trans. Ind. Electron.* 69, 10510–10520. doi: 10.1109/TIE.2022.3161810

Suykens, J., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural. Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Tang, H, and Jia, K, Discriminative adversarial domain adaptation. In: Proceeding of the 34th National Conference on artificial intelligence, Feb. 7-12, New York (2019).

Tao, J., Chung, F. L., and Wang, S. (2012). On minimum distribution discrepancy support vector machine for domain adaptation. *Pattern Recogn.* 45, 3962–3984. doi: 10.1016/j.patcog.2012.04.014

Tao, J. W., and Dan, Y. F. (2021). Multi-source co-adaptation for EEG-based emotion recognition by mining correlation information. *Front. Neurosci.* 15:677106. doi: 10.3389/fnins.2021.677106

Tao, J., Dan, Y., and Di, Z. (2021). Robust multi-source co-adaptation with adaptive loss minimization. *Signal Process. Image Commun.* 99:116455. doi: 10.1016/j.image.2021.116455

Tao, J., Dan, Y. F., Zhou, D., and He, S. S. (2022). Robust latent multi-source adaptation for encephalogram-based emotion recognition. *Front. Neurosci.* 16:850906. doi: 10.3389/fnins.2022.850906

Tao, J., Di Zhou, F. L., and Zhu, B. (2019). Latent multi-feature co-regression for visual recognition by discriminatively leveraging multi-source models. *Pattern Recogn.* 87, 296–316. doi: 10.1016/j.patcog.2018.10.023

Tao, J. W., Song, D., Wen, S., and Hu, W. (2017). Robust multi-source adaptation visual classification using supervised low-rank representation. *Pattern Recogn.* 61, 47–65. doi: 10.1016/j.patcog.2016.07.006

Tao, J., Wen, S., and Hu, W. (2015). L1-norm locally linear representation regularization multi-source adaptation learning. *Neural Netw.* 69, 80–98. doi: 10.1016/j.neunet.2015.01.009

Tao, J., Wen, S., and Hu, W. (2016). Multi-source adaptation learning with global and local regularization by exploiting joint kernel sparse representation. *Knowl. Based Syst.* 98, 76–94. doi: 10.1016/j.knosys.2016.01.021

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: maximizing for domain invariance. *CoRR* abs/1412.3474 Available at: http://arxiv.org/abs/1412.3474

Wang, J., Ji, Z., Kim, H. E., Wang, S., Xiong, L., and Jiang, X. (2017). Selecting optimal subset to release under differentially private M-estimators from hybrid datasets. *IEEE Trans. Knowl. Data Eng.* 30, 573–584. doi: 10.1109/TKDE.2017.2773545

Xiao, X., Xu, M., Jin, J., Wang, Y., Jung, T. P., and Ming, D. (2020). Discriminative canonical pattern matching for single-trial classification of erp components. *IEEE Trans. Biomed. Eng.* 67, 2266–2275. doi: 10.1109/TBME.2019.2958641

Zhang, Y., Dong, J., Zhu, J., and Wu, C. (2019b). Common and special knowledge-driven TSK fuzzy system and its modeling and application for epileptic EEG signals recognition. *IEEE Access*, 2019 7, 127600–127614. doi: 10.1109/ACCESS.2019.2937657

Zhang, Y., Tian, F., Wu, H., Geng, X., Qian, D., Dong, J., et al. (2017). Brain MRI tissue classification based fuzzy clustering with competitive learning. *J. Med. Imaging Health Informat.* 7, 1654–1659. doi: 10.1166/jmihi.2017.2181

Zhang, Y., Wang, S., Xia, K., Jiang, Y., and Qian, P. (2021). Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Informat. Fusion* 66, 170–183. doi: 10.1016/j.inffus.2020.09.002

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2019a). Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* 49, 839–847. doi: 10.1109/TCYB.2017.2788081

Zheng, W. (2017). Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis. *IEEE Trans. Cogn. Dev. Syst.* 9, 281–290. doi: 10.1109/TCDS.2016.2587290

Zheng, W.-L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2019). EmotionMeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W.-L., and Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zheng, W. L., and Lu, B. L. *Personalizing EEG-based affective models with transfer learning*. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press (2016), pp. 2732–2738.

Zheng, W. L., Zhang, Y. Q., Zhu, J. Y., and Lu, B. L. (2015). "Transfer components between subjects for EEG-based emotion recognition" in *International conference on affective computing and intelligent interaction (ACII)* (Xi'an), 917–922.

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* doi: 10.48550/arXiv.1907.07835

Zhou, R., Zhang, Z., Fu, H., Zhang, L., Li, L., Huang, G., et al. (2022). A novel transfer learning framework with prototypical representation based pairwise learning for cross-subject cross-session EEG-based emotion recognition. *ArXiv abs/2202.06509*. doi: 10.48550/arXiv.2202.06509

Zhu, J., Arbor, A., and Hastie, T. (2006). Multi-class adaboost. *Stat. Interface* 2, 349–360. doi: 10.4310/SII.2009.v2.n3.a8

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 04 January 2024
DOI 10.3389/fnins.2023.1281809

OPEN ACCESS

EDITED BY
Xi Jiang,
University of Electronic Science and Technology of China, China

REVIEWED BY
Man Fai Leung,
Anglia Ruskin University, United Kingdom
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China

*CORRESPONDENCE
Gang Xiao
✉ xiao.math@foxmail.com
Xiaolei Zhang
✉ bmezhang@vip.163.com

†These authors share first authorship

RECEIVED 23 August 2023
ACCEPTED 20 November 2023
PUBLISHED 04 January 2024

CITATION
Yang W, Zou J, Zhang X, Chen Y, Tang H, Xiao G and Zhang X (2024) An end-to-end LSTM-Attention based framework for quasi-steady-state CEST prediction. *Front. Neurosci.* 17:1281809. doi: 10.3389/fnins.2023.1281809

An end-to-end LSTM-Attention based framework for quasi-steady-state CEST prediction

Wei Yang[1,2†], Jisheng Zou[2†], Xuan Zhang[2†], Yaowen Chen[2], Hanjing Tang[2], Gang Xiao[3*] and Xiaolei Zhang[4*]

[1]Great Bay University, Dongguan, China, [2]College of Engineering, Shantou University, Shantou, China, [3]School of Mathematics and Statistics, Hanshan Normal University, Chaozhou, China, [4]Department of Radiology, Second Affiliated Hospital of Shantou University Medical College, Shantou, China

Chemical exchange saturation transfer (CEST)-magnetic resonance imaging (MRI) often takes prolonged saturation duration (Ts) and relaxation delay (Td) to reach the steady state, and yet the insufficiently long Ts and Td in actual experiments may underestimate the CEST measurement. In this study, we aimed to develop a deep learning-based model for quasi-steady-state (QUASS) prediction from non-steady-state CEST acquired in experiments, therefore overcoming the limitation of the CEST effect which needs prolonged saturation time to reach a steady state. To support network training, a multi-pool Bloch-McConnell equation was designed to derive wide-ranging simulated Z-spectra, so as to solve the problem of time and labor consumption in manual annotation work. Following this, we formulated a hybrid architecture of long short-term memory (LSTM)-Attention to improve the predictive ability. The multilayer perceptron, recurrent neural network, LSTM, gated recurrent unit, BiLSTM, and LSTM-Attention were included in comparative experiments of QUASS CEST prediction, and the best performance was obtained by the proposed LSTM-Attention model. In terms of the linear regression analysis, structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and mean-square error (MSE), the results of LSTM-Attention demonstrate that the coefficient of determination in the linear regression analysis was at least $R^2 = 0.9748$ for six different representative frequency offsets, the mean values of prediction accuracies in terms of SSIM, PSNR and MSE were 0.9991, 49.6714, and 1.68×10^{-4} for all frequency offsets. It was concluded that the LSTM-Attention model enabled high-quality QUASS CEST prediction.

KEYWORDS

CEST-MRI, QUASS CEST, deep learning, Bloch-McConnell equation, LSTM-Attention

1 Introduction

Chemical exchange saturation transfer (CEST)-magnetic resonance imaging (MRI) of dilute labile protons that undergo their chemical exchange with the bulk water protons enables a specific contrast and provides a promising molecular imaging tool for *in vivo* applications (Zaiss and Bachert, 2013; Xiao et al., 2015; Wu et al., 2016; Jones et al., 2018; Zaiss et al., 2022). However, the CEST effect is limited by experimental conditions such as the amplitude (Sun et al., 2007; Zhao et al., 2011) and duration of RF saturation (Randtke et al., 2014; Zaiss et al., 2018). For some CEST-MRI experiments, the CEST effect needs

prolonged saturation duration to achieve quasi-steady-state (QUASS). The limitation of maximum RF saturation duration underestimates the CEST signal (Zhang et al., 2021), which makes it difficult to compare the results between different platforms and stations (Sun, 2021; Wu et al., 2022). So the task for a post-processing strategy to automatically derive the QUASS CEST effect from experimental measurements with limited saturation duration needs to be solved today. Particularly, Sun conducted a QUASS CEST analysis that compensated the effect of finite saturation duration (Ts) and relaxation delay (Td) by solving both the labile proton fraction ratio and exchange rate from simulated CEST, therefore improving the accuracy of CEST-MRI quantification (Sun, 2021). Zhang et al. developed a postprocessing strategy to derive the QUASS CEST by modeling the CEST signal evolution as a function of Ts and Td, allowing robust CEST quantification (Zhang et al., 2021). Kim et al. proposed a QUASS CEST algorithm that can minimize dependences on Ts and Td by combining multi-slice CEST imaging with QUASS processing (Kim et al., 2022).

The application of deep learning to the CEST-MRI has led to a large number of technical improvements (Glang et al., 2020; Li et al., 2020; Bie et al., 2022; Huang et al., 2022; Perlman et al., 2022), including shortcut of the conventional Lorentzian fitting for *in vivo* 3 T CEST data (Glang et al., 2020), prediction of the CEST contrasts for Alzheimer's disease (Huang et al., 2022), identification of pertinent Z-spectral features for distinguishing tumor aggressiveness (Bie et al., 2022), etc. Therefore, this paper aims to employ a deep learning technique to predict QUASS CEST (i.e., CEST images on prolonged saturation) by training a network on the prior knowledge of simulated CEST Z-spectra. With respect to the underlying application domain, the sequence-to-sequence (Seq2Seq) network is an intuitive approach, in which the LSTM (Yu et al., 2019) and the attention mechanism (Vaswani et al., 2017) are stated as excellent methods. LSTM is known to solve the vanishing gradient problem when a recurrent neural network (RNN) is used to work with the sequence input, while the disadvantage of LSTM in the latent decomposition training is significant (Shi et al., 2022). For this, some modified versions help improve the LSTM performance and were successfully applied to medical treatment behavior prediction (Cheng et al., 2021), medical event prediction (Liu et al., 2022), and EEG-based emotion recognition (Chakravarthi et al., 2022). Particularly, a Seq2Seq with a multi-head attention mechanism instead of recurrence has excelled at tasks of time series, obtaining effective information and significant spatiotemporal features from the new coding sequence. However, the attention mechanism loses the sequential information because the used attention mechanism is position-insensitive (Zheng et al., 2021). The principles of the LSTM network and attention mechanism were briefly described in Supplementary File 1. Looking at the advantages and disadvantages of both algorithms, a hybrid model named LSTM-Attention would be a perfectly natural way. In this LSTM-Attention architecture, the LSTM is used to obtain the hidden state of the input features, while the use of multi-head attention in the encoder layer is to better learn the temporal information (Figure 1). In Figure 1, the input embedding is used to capture high-dimensional spatial properties of long time series. Because position information is not considered in the attention layer, we add "positional encoding" to the input embeddings. To this end, different semantic information from different sequence positions is incorporated into an embedding tensor, compensating for the lack of position information. The LSTM had a hidden state dimensionality of 1,024, the number of attention heads is 4.

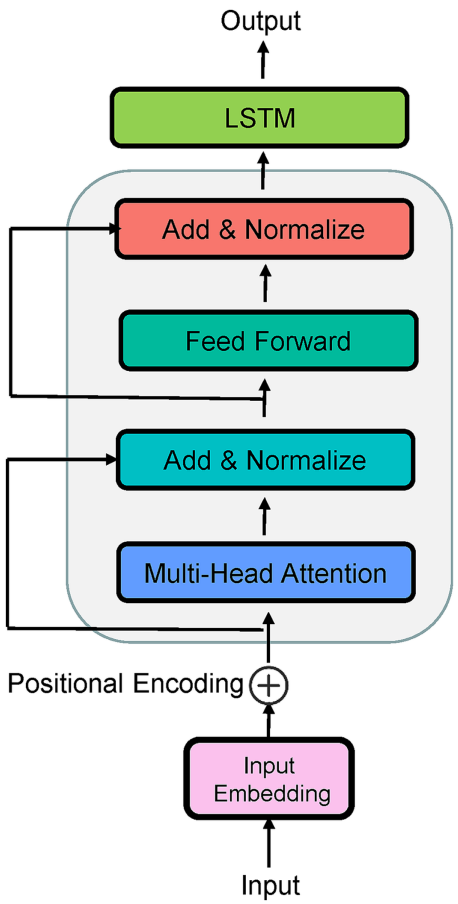


FIGURE 1
Structure diagram of LSTM-Attention.

Motivated by the above, this paper aims to build an LSTM-Attention-based model for QUASS CEST prediction from non-steady-state CEST (i.e., CEST images with shorter saturation time) acquired in experiments, as shown in Figure 2. Simulated Z-spectra with shorter and prolonged saturation time was derived from the designed Bloch–McConnell equations (Xiao et al., 2023), respectively. Then we used the trained model to predict QUASS CEST from non-steady-state CEST acquired in experiments.

In summary, this work makes the following two key contributions. To tackle the problematic and time-consuming task of obtaining the labeled training data from experiments, we built a large-scale training set based on simulated Z-spectra derived from the designed Bloch-McConnell equations. We formulated an LSTM-Attention-based model which is trained on simulated CEST Z-spectra to predict QUASS CEST image pixel-by-pixel from non-steady-state CEST acquired in explements, where the attention mechanism improves the predictive ability of LSTM by paying attention to the input weights that contribute more to the output.

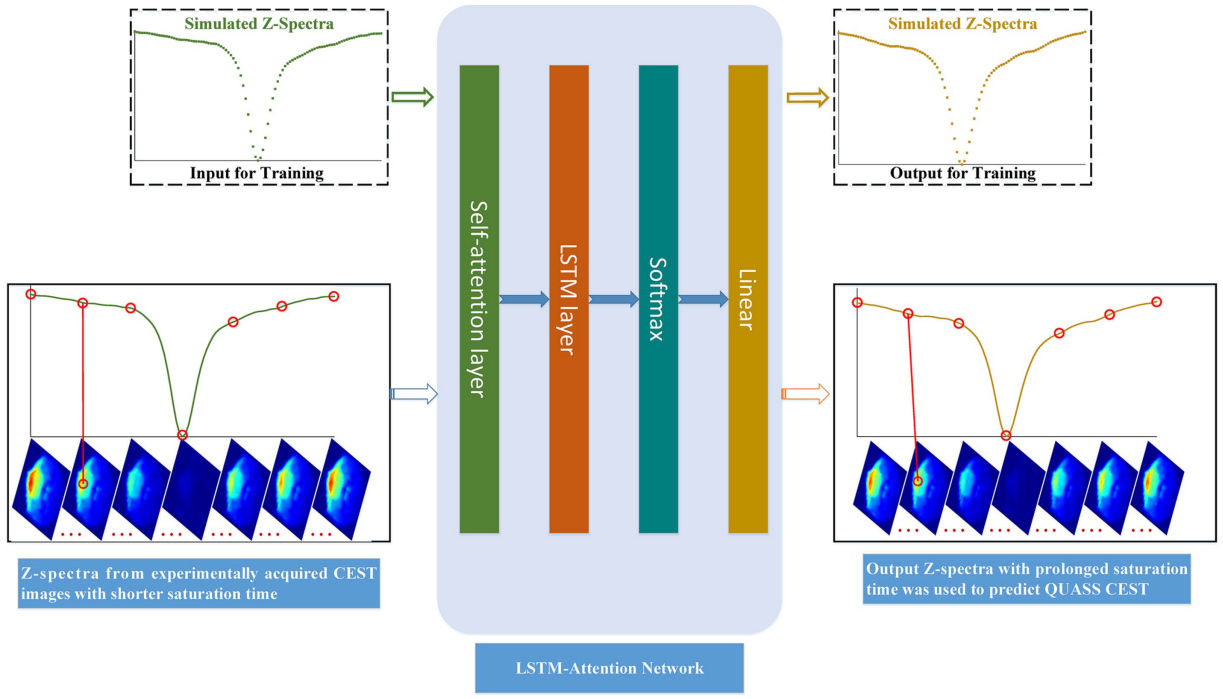


FIGURE 2
Flow chart of LSTM-Attention based model for predicting QUASS CEST. The simulated Z-spectra with short saturation time was referred to the training input, and the simulated Z-spectra with prolonged saturation time was the training output. Input Z-spectra from experiments with shorter saturation times for each image pixel, we can predict QUASS CEST images by their output Z-spectra of each pixel.

2 Materials and methods

2.1 *In vivo* MRI experiments

In this vivo MRI experiment, 8-week-old male SD rats (Beijing Vital River Laboratory Animal Technology Co., Ltd.) weighing 250 g were used to generate a tumor-bearing model. All animal care and experimental procedures were performed in accordance with the National Research Council Guide for the Care and Use of Laboratory Animals. For this assessment, a 10 μL suspension of rat glioma C6 cells (approximately 2 × 106 cells) was implanted into the right basal ganglia (specific injection position: AP + 1, ML + 3, DV-5) of the rats using a Hamilton syringe and a 30-gauge needle. Two weeks after the implantation of tumor cells, the rats were subjected to MRI.

The CEST-MRI experiment was performed using a 7 T horizontal bore small animal MRI scanner (Agilent Technologies, Santa Clara, CA, U.S.A.) with a surface coil (Timemedical Technologies, China) for transmission and reception. Imaging parameters were as follows: repetition time (TR) = 6,000 ms, echo time (TE) = 40 ms, array = frequency offsets, slice thickness = 2 mm, field of view (FOV) = 64 × 64 mm, matrix size = 64 × 64, spatial resolution = 1 × 1 mm, averages = 1. To obtain CEST images, an echo planar imaging readout sequence was used, where continuous wave (CW) RF irradiation was implemented on scanners. The saturation times were 1.5 s and 5 s, respectively, with 101 frequency offsets evenly distributed from −6 to 6 ppm relative to the resonance of water.

The CEST images of saturation times 1.5 s acquired in this experiment were the inputs of trained networks. The CEST images with saturation times 5 s acquired in this experiment were the reference, which is used to assess the prediction performance by comparing the model's estimates with the experimental data values.

2.2 Training dataset

The training of LSTM-Attention for predicting objects requires a large dataset with true pixel-level labels in terms of saturation times, which is extremely expensive to construct training data in experiments. To address this issue, we simulated CEST signals using a 7-pool Bloch–McConnell equation (Xiao et al., 2023) at both non-steady and quasi-steady states. This 7-pool model consists of free water centered at 0 ppm, amide centered at 3.5 ppm, guanidyl/amine centered at 2.0 ppm, hydroxyl centered at 1.3 ppm, nuclear Overhauser enhancement (NOE) centered at-1.6 ppm, magnetization transfer (MT) centered at-2.4 ppm, and NOE centered at-3.5 ppm. In detail, 20 dynamic parameters regarding all possible tissue combinations were considered. For each dynamic parameter, random variables from the uniform distribution with lower bound and upper bound were sampled for the training dataset, so we could generate as much data as needed with all possible tissue combinations. The sampled variables of each parameter interacting with that of each other generated 350,000 parameter combinations, thus yielding 350,000 paired simulated Z-spectra (see Supplementary Figure S1). The simulated Z-spectra with saturation times of 1.5 s and 5 s at 101 offsets in the range of ±6 ppm were referred to as the training input and output, respectively.

2.3 Evaluation metrics and workstation

Linear regression analysis (Li et al., 2020) was first applied to evaluate the proposed model at frequency offsets-3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm. To evaluate the proposed model in the prediction of CEST image at each frequency offset, the prediction performance was evaluated by three measures: the structural similarity index (SSIM), the peak signal-to-noise ratio (PSNR) (Hore and Ziou, 2010), and mean squared error (MSE).

The workstation used in this study is a Lenovo ST558 workstation with 32 G memory, a dual-core CPU10 core, and a 2.4 G main operating frequency. The experiments are based on PyTorch, and the number of epochs is 100. The number of batch size is 256. The optimizer is Adam, and the learning rate is 0.0001. We initialize the weights using samples from a uniform distribution, and use MSE-Loss as the loss function.

3 Results

To validate the proposed model, prediction images were compared with the reference from experimental measurements. We applied the trained neural networks to predict the CEST images with a saturation time of 5 s from experimentally acquired CEST images with a saturation time of 1.5 s. For comparison, the LSTM-Attention presented comparable performance to that of five popular existing networks: the multilayer perceptron (MLP) (Xu et al., 2018), recurrent neural network (RNN) (Xu et al., 2018), long short-term memory (LSTM) (Yu et al., 2019), gated recurrent unit (GRU) (Xu et al., 2018), and BiLSTM (Siami-Namini et al., 2019).

We first conducted an experiment to predict CEST images at frequency offsets-3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm, as shown in Figure 3. The region of the pseudo color image overlaid on the anatomy image was the region of interest (ROI). The results obtained from the considered networks were almost equivalent to those obtained experimentally by the subjective vision.

Furthermore, we carried out a comparison experiment in terms of the absolute error modulus between reference and prediction, as illustrated in Figure 4. In this figure, row-plots indicated the absolute error modulus at frequency offsets −3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm; columns (A–F) were the absolute error modulus from MLP, RNN, LSTM, GRU, BiLSTM and LSTM-Attention, respectively; the plot (G) denoted the mean values of

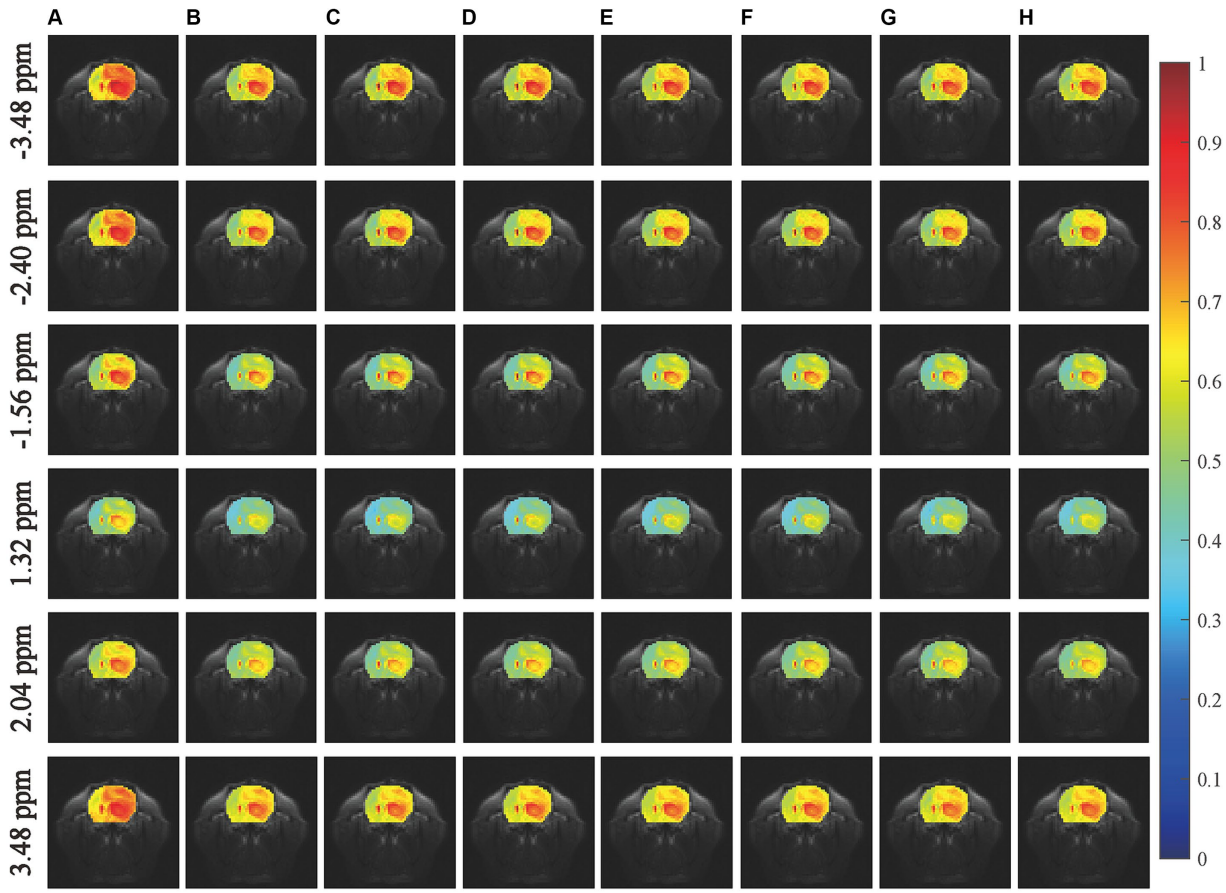


FIGURE 3
Comparisons of the predicted results with experimentally acquired CEST images at frequency offsets-3.48 ppm, −2.40 ppm, −1.56 ppm, 0.96 ppm, 2.04 ppm, and 3.48 ppm. The column **(A)** shows the experimentally acquired CEST image with the saturation time of 1.5 s, the column **(B)** shows the experimentally acquired CEST image with the saturation time of 5 s (reference), the columns **(C–H)** denote the prediction results obtained by MLP, RNN, LSTM, GRU, BiLSTM and LSTM-Attention, respectively.

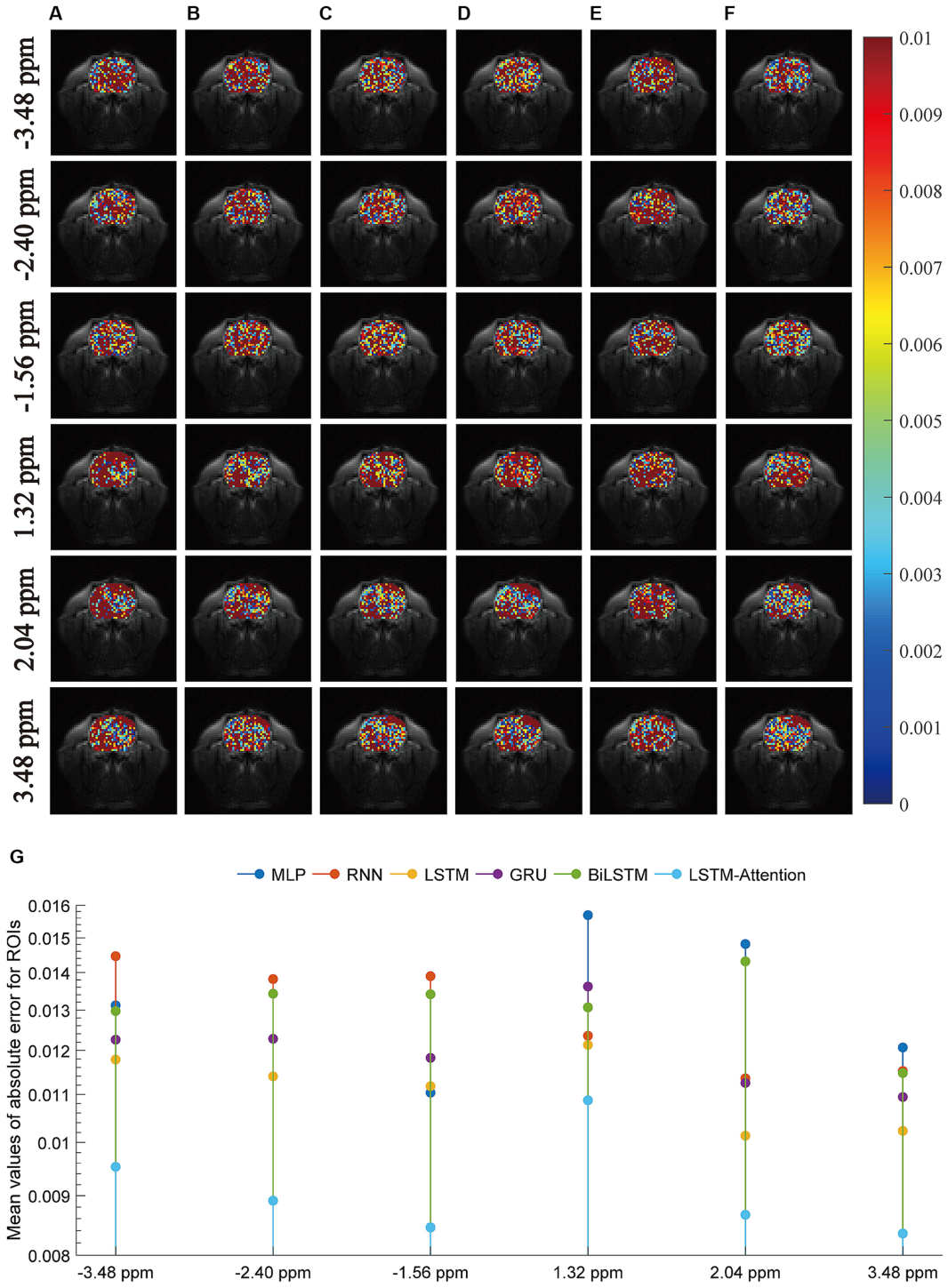


FIGURE 4
The absolute error modulus between the predicted images and the experimentally acquired CEST images at frequency offsets-3.48 ppm, −2.28 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm. The columns **(A–F)** are the results from MLP, RNN, LSTM, GRU, BiLSTM and LSTM-Attention, respectively; the plot **(G)** denotes the mean values of absolute error modulus obtained by considered six networks.

absolute error modulus at frequency offsets −3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm and 3.48 ppm that obtained by considered methods. The results of Figure 4G reveal that the mean values of absolute error modulus obtained from the proposed LSTM-Attention model are smaller than those of other networks at these frequency offsets, while there is a difference of one order of magnitude between the LSTM-Attention and its counterparts at frequency offsets-3.48 ppm, −2.40 ppm, −1.56 ppm, 2.04 ppm, and 3.48 ppm. In other words, the CEST image at these frequency offsets obtained by the trained LSTM-Attention showed a higher degree of agreement with those obtained by the experimental measurements as the standard.

An example of the predicted Z-spectra by LSTM-Attention for white matter, gray matter, and tumor is shown in Figure 5, which consistently provided satisfactory results.

Figure 6 quantitatively demonstrates the considered networks for predicting the *in vivo* CEST signal by plotting the linear regression lines and scatter diagrams between the reference and the prediction. In this figure, row-plots were the results at frequency offsets −3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm; columns (A–F) denoted the results from MLP, RNN, LSTM, GRU, BiLSTM and LSTM-Attention, respectively; the plot (G) denoted the coefficient of determination at frequency offsets −3.48 ppm, −2.40 ppm, −1.56 ppm, 1.32 ppm, 2.04 ppm, and 3.48 ppm that obtained by considered methods. The pixel values correspond to the points of the ROI in Figures 3, 4. For each plot, the fitting curve was denoted by the blue line and the green line was the 45-degree diagonal. The excellent performance of our prediction was confirmed by the scatter and linear regression lines, resulting in a very high coefficient of determination ($R^2 \geq 0.9748$) at these frequency offsets.

To set up a comprehensive way to evaluate the performance of the prediction models, the SSIM and PSNR from the reference and the prediction at each offset (−6 ~ 6 ppm) are considered, as displayed in Figure 7. In terms of SSIM, the LSTM-Attention exhibits good accuracies at each offset (−6 ~ 6 ppm) and presents results close to those of LSTM at-5.04 and 0.96 ppm, while it exceeds the performance of other networks in the ranges (−6 ~ −5.16 ppm), (−4.92 ~ −0.72 ppm) and (1.08 ~ 4.68 ppm). Similar results are obtained by LSTM-Attention in terms of PSNR. Clearly, our model exhibits competitive results for these two metrics based on different criteria, providing a mean SSIM value of 0.9991 and a mean PSNR value of 49.6714, respectively. Figure 8 displays the MSE obtained by considered networks for all frequency offsets, and the best result of mean MSE 1.68×10^{-4} is obtained by the LSTM-Attention network.

4 Discussions

To some extent, we developed a general deep learning-based approach to predict QUASS CEST using experimentally acquired CEST images with shorter saturation times, since the performances of MLP and five existing Seq2Seq networks are also evaluated in this study. As the results show, the LSTM-Attention network outperforms the MLP, RNN, LSTM, GRU, and BiLSTM (Figure 7). It is clear that LSTM-Attention is able to capture the underlying context better by paying attention to the input weights that contribute more to the output. The better performance of LSTM-Attention compared to its counterparts is understandable for certain types of data such as specific chemical groups in the downfield and MT/NOE in the upfield (Figures 4, 6).

In fact, the Z-spectra of a pixel typically behaves short-and long-range dependencies along the frequency offsets (see Supplementary Figure S2). The LSTM-Attention is consistently the best model followed by MLP, RNN, LSTM, GRU, and BiLSTM for capturing the short-and long-range behavior. In the simplest form, fully RNN is an MLP with the previous set of hidden unit activations feeding back into the network along with the inputs (Roy et al., 2019). Additionally, the LSTM, GRU, BiLSTM, and LSTM-Attention are able to overcome RNN's vanishing gradient problem which happens when RNN learns long-range dependencies of inputs (Yang et al., 2022). Therefore, the ability of short-and long-range interaction in these considered networks performs similarly, as the results above. Particularly, the LSTM-Attention augments the non-linear processing

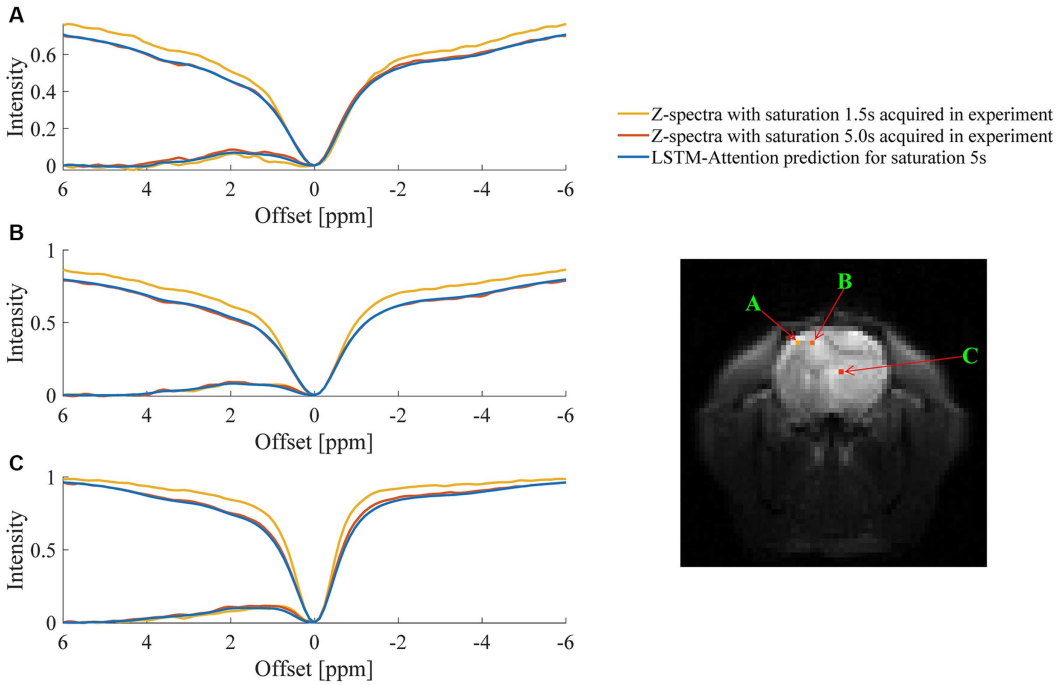


FIGURE 5
Comparison between the predicted Z-spectra of LSTM-Attention and the experimentally acquired results at one randomly chosen pixel of **(A)** gray matter, **(B)** white matter and **(C)** tumor, respectively.

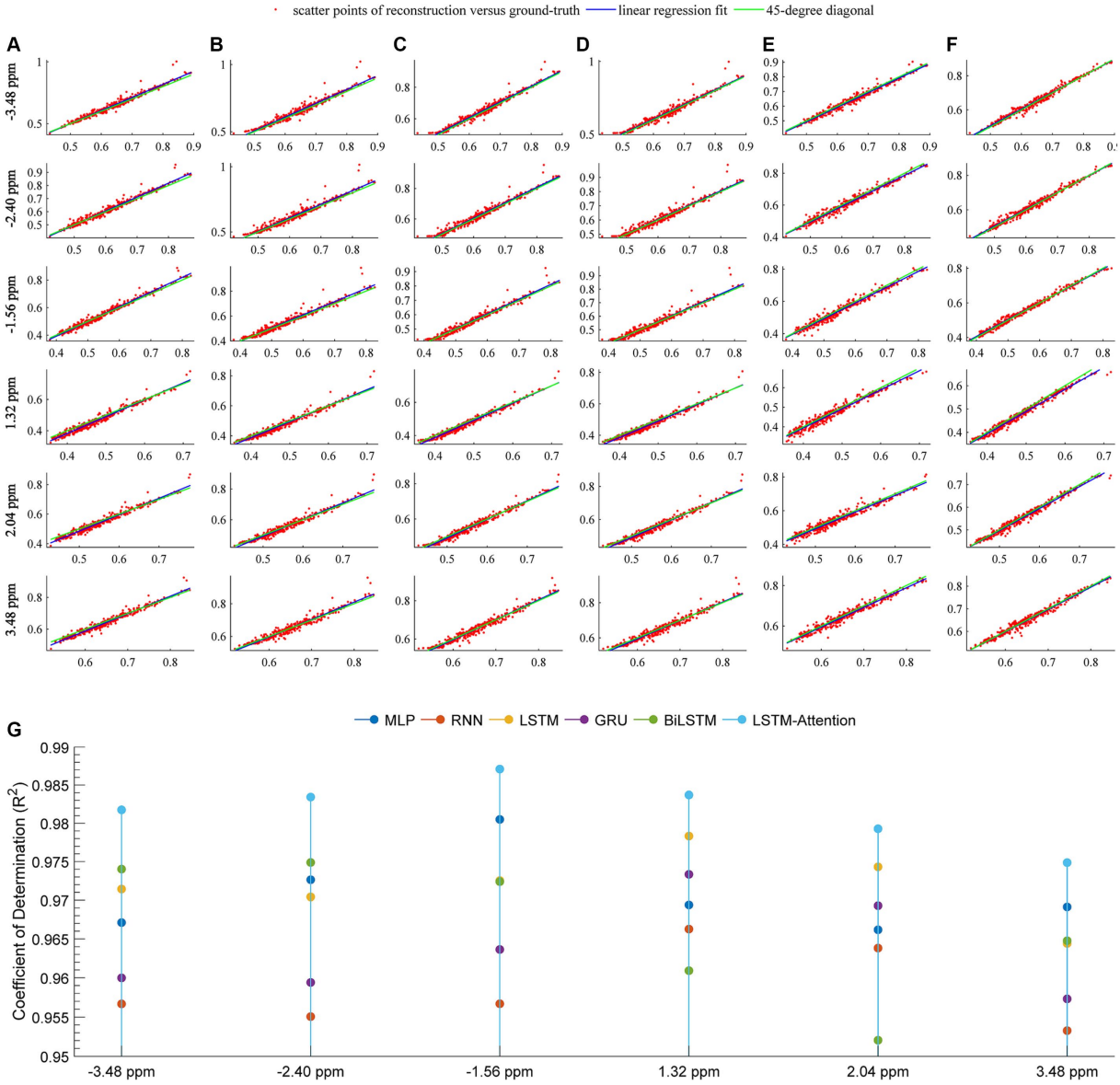


FIGURE 6
Linear regression analysis of the prediction and reference at frequency offsets-3.48, −2.40, −1.56, 1.32, 2.04, 3.48ppm. The columns **(A–F)** are the results from MLP, RNN, LSTM, GRU, BiLSTM and LSTM-Attention, respectively; the plot **(G)** denotes the coefficient of determination (R^2) between the prediction and reference. For the columns **(A–F)** at each offset, the locations of the red markers are specified by the vectors *x* and *y*, where *x* is the pixel values of the experimentally acquired CEST image with the saturation time 5s (reference) and y is the pixel values of predicted CEST images with a saturation time 5s; the blue line is the linear regression fitting based on the red scatter points of prediction versus reference, the green line indicates the 45-degree diagonal.

capability in QUASS CEST prediction by taking advantage of the known, observed, and static covariate factors.

In practice, training a deep neural network to predict QUASS CEST requires massive samples with ground-truth annotations, which is extremely expensive to construct experimentally. To solve this problem, we built an automatically labeled dataset based on the Bloch–McConnell equations. Briefly, we considered all the possible parameters of the equations when generating the trained samples. For each dynamic parameter, a wide range of random values was sampled in a uniform distribution with its lower and upper bounds, automatically yielding a large set of labeled training data.

Further studies would be beneficial for QUASS CEST applications at low-field MRI where short saturation time is needed. It could be useful to investigate other less visible CEST effects (such as guanidyl or amine) in clinical MRI scanners.

5 Conclusion

In summary, we addressed the QUASS CEST predicting problem in learning systems and proposed a data-driven predicting scheme that benefits from our strategy to reduce the effect of finite RF saturation duration on the CEST measurement. The experiment study

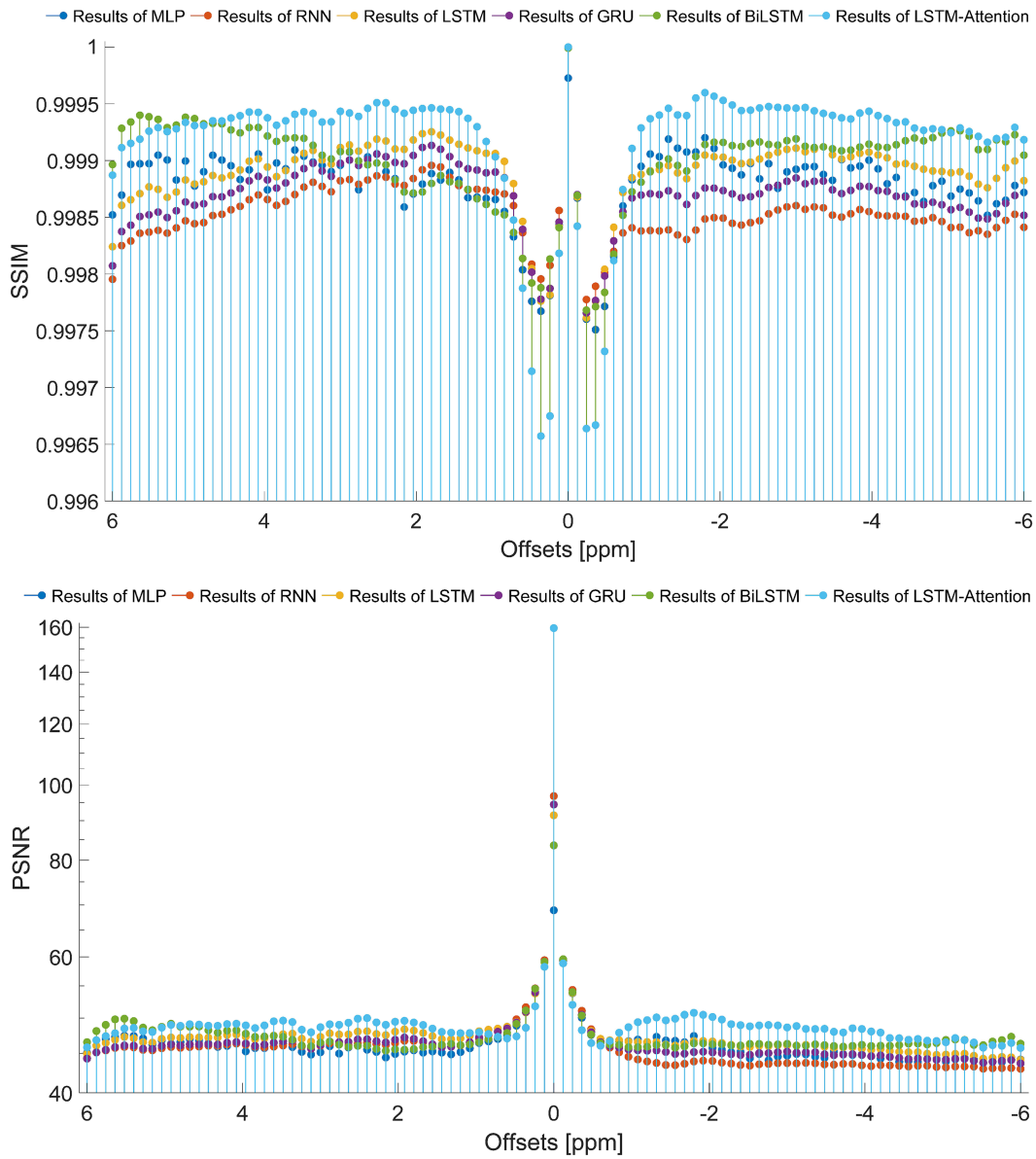


FIGURE 7
The SSIM and PSNR obtained from the reference and the prediction at each offset (–6 ~ 6 ppm).

compared the proposed model with other approaches, and the effectiveness and superiority of the LSTM-Attention model were validated. This research can be further expanded to predict problems for available clinical MRI scanners.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The animal study was approved by National Research Council Guide for the Care and Use of Laboratory Animals. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

WY: Formal analysis, Methodology, Software, Visualization, Writing – original draft. JZ: Data curation, Formal analysis, Software, Visualization, Writing – original draft. XuZ: Data curation, Formal analysis, Software, Visualization, Writing – original draft. YC: Funding acquisition, Investigation, Project administration, Resources, Writing – review & editing. HT: Data curation, Resources, Visualization, Writing –review & editing. GX: Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – review & editing. XiZ: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing.

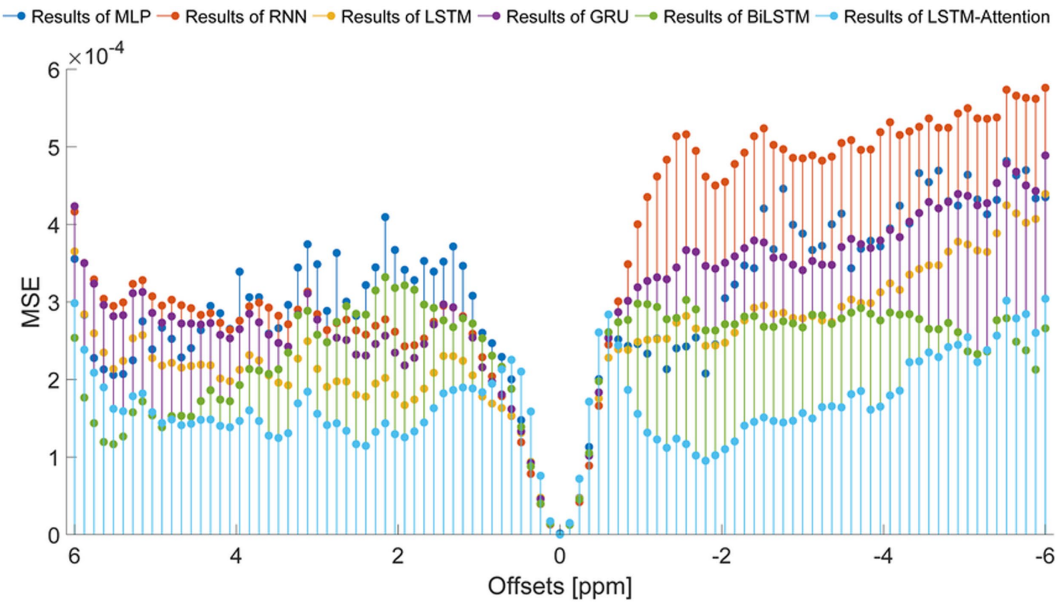


FIGURE 8
The MSE obtained from the reference and the prediction at each offset (–6 ~ 6 ppm).

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study was supported by the 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grants (Grant/Award Numbers: 2020LKSFG06C) and the Medical Health Science and Technology Project of Shantou (Grant/Award Number: 2022-88-16).

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1281809/full#supplementary-material

References

Bie, C., Li, Y., Zhou, Y., Bhujwalla, Z. M., Song, X., Liu, G., et al. (2022). Deep learning-based classification of preclinical breast cancer tumor models using chemical exchange saturation transfer magnetic resonance imaging. *NMR Biomed.* 35:e4626. doi: 10.1002/nbm.4626

Chakravarthi, B., Ng, S.-C., Ezilarasan, M., and Leung, M.-F. (2022). EEG-based emotion recognition using hybrid CNN and LSTM classification. *Front. Comput. Neurosci.* 16:1019776. doi: 10.3389/fncom.2022.1019776

Cheng, L., Shi, Y., Zhang, K., Wang, X., and Chen, Z. (2021). GGATB-LSTM: grouping and global attention-based time-aware bidirectional LSTM medical treatment behavior prediction. *ACM Trans Knowl Discov Data* 15, 1–16. doi: 10.1145/3441454

Glang, F., Deshmane, A., Prokudin, S., Martin, F., Herz, K., Lindig, T., et al. (2020). DeepCEST 3T: robust MRI parameter determination and uncertainty quantification with neural networks—application to CEST imaging of the human brain at 3T. *Magn. Reson. Med.* 84, 450–466. doi: 10.1002/mrm.28117

Hore, A., and Ziou, D. (2010). "Image quality metrics: PSNR vs. SSIM." *2010 20th international conference on pattern recognition*, IEEE.

Huang, J., Lai, J. H., Tse, K. H., Cheng, G. W., Liu, Y., Chen, Z., et al. (2022). Deep neural network based CEST and AREX processing: application in imaging a model of Alzheimer's disease at 3 T. *Magn. Reson. Med.* 87, 1529–1545. doi: 10.1002/mrm.29044

Jones, K. M., Pollard, A. C., and Pagel, M. D. (2018). Clinical applications of chemical exchange saturation transfer (CEST) MRI. *J. Magn. Reson. Imaging* 47, 11–27. doi: 10.1002/jmri.25838

Kim, H., Krishnamurthy, L. C., and Sun, P. Z. (2022). Demonstration of fast multi-slice quasi-steady-state chemical exchange saturation transfer (QUASS CEST) human brain imaging at 3T. *Magn. Reson. Med.* 87, 810–819. doi: 10.1002/mrm.29028

Li, Y., Xie, D., Cember, A., Nanga, R. P. R., Yang, H., Kumar, D., et al. (2020). Accelerating GluCEST imaging using deep learning for B0 correction. *Magn. Reson. Med.* 84, 1724–1733. doi: 10.1002/mrm.28289

Liu, S., Wang, X., Xiang, Y., Xu, H., Wang, H., and Tang, B. (2022). Multi-channel fusion LSTM for medical event prediction using EHRs. *J. Biomed. Inform.* 127:104011. doi: 10.1016/j.jbi.2022.104011

Perlman, O., Zhu, B., Zaiss, M., Rosen, M. S., and Farrar, C. T. (2022). An end-to-end AI-based framework for automated discovery of rapid CEST/MT MRI acquisition protocols and molecular parameter quantification (AutoCEST). *Magn. Reson. Med.* 87, 2792–2810. doi: 10.1002/mrm.29173

Randtke, E. A., Chen, L. Q., and Pagel, M. D. (2014). The reciprocal linear QUEST analysis method facilitates the measurements of chemical exchange rates with CEST MRI. *Contrast Media Mol. Imaging* 9, 252–258. doi: 10.1002/cmmi.1566

Roy, K., Mandal, K. K., and Mandal, A. C. (2019). Ant-lion optimizer algorithm and recurrent neural network for energy management of micro grid connected system. *Energy* 167, 402–416. doi: 10.1016/j.energy.2018.10.153

Shi, H., Gao, S., Tian, Y., Chen, X., and Zhao, J. (2022). "Learning bounded context-free-grammar via LSTM and the transformer: difference and the explanations," in *Proceedings of the AAAI conference on artificial intelligence*. Palo Alto, California: AAAI Press, 8267–8276.

Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). "The performance of LSTM and BiLSTM in forecasting time series." in *2019 IEEE international conference on big data*, IEEE.

Sun, P. Z. (2021). Quasi-steady-state chemical exchange saturation transfer (QUASS CEST) solution improves the accuracy of CEST quantification–QUASS CEST MRI-based omega plot analysis. *Magn. Reson. Med.* 86, 765–776. doi: 10.1002/mrm.28744

Sun, P. Z., Zhou, J., Huang, J., and Van Zijl, P. (2007). Simplified quantitative description of amide proton transfer (APT) imaging during acute ischemia. *Magn Reson Med* 57, 405–410. doi: 10.1002/mrm.21151

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30. doi: 10.48550/arXiv.1706.03762

Wu, Y., Liu, Z., Yang, Q., Zou, L., Zhang, F., Qian, L., et al. (2022). Fast and equilibrium CEST imaging of brain tumor patients at 3T. *Neuroimage Clin* 33:102890. doi: 10.1016/j.nicl.2021.102890

Wu, B., Warnock, G., Zaiss, M., Lin, C., Chen, M., Zhou, Z., et al. (2016). An overview of CEST MRI for non-MR physicists. *EJNMMI Phys* 3, 1–21. doi: 10.1186/s40658-016-0155-2

Xiao, G., Sun, P. Z., and Wu, R. (2015). Fast simulation and optimization of pulse-train chemical exchange saturation transfer (CEST) imaging. *Phys. Med. Biol.* 60, 4719–4730. doi: 10.1088/0031-9155/60/12/4719

Xiao, G., Zhang, X., Yang, G., Jia, Y., Yan, G., and Wu, R. (2023). Deep learning to reconstruct quasi-steady-state chemical exchange saturation transfer from a non-steady-state experiment. *NMR Biomed.* 36:e4940. doi: 10.1002/nbm.4940

Xu, C., Shen, J., Du, X., and Zhang, F. (2018). An intrusion detection system using a deep neural network with gated recurrent units. *IEEE Access* 6, 48697–48707. doi: 10.1109/ACCESS.2018.2867564

Yang, J., Chen, X., Wang, D., Zou, H., Lu, C. X., Sun, S., et al. (2022). Deep learning and its applications to wifi human sensing: a benchmark and a tutorial. arXiv preprint arXiv:2207.07859.

Yu, Y., Si, X., Hu, C., and Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* 31, 1235–1270. doi: 10.1162/neco_a_01199

Zaiss, M., Angelovski, G., Demetriou, E., McMahon, M. T., Golay, X., and Scheffler, K. (2018). QUESP and QUEST revisited–fast and accurate quantitative CEST experiments. *Magn. Reson. Med.* 79, 1708–1721. doi: 10.1002/mrm.26813

Zaiss, M., and Bachert, P. (2013). Chemical exchange saturation transfer (CEST) and MR Z-spectroscopy in vivo: a review of theoretical approaches and methods. *Phys. Med. Biol.* 58, R221–R269. doi: 10.1088/0031-9155/58/22/R221

Zaiss, M., Jin, T., Kim, S. G., and Gochberg, D. F. (2022). Theory of chemical exchange saturation transfer MRI in the context of different magnetic fields. *NMR Biomed.* 35:e4789. doi: 10.1002/nbm.4789

Zhang, X. Y., Zhai, Y., Jin, Z., Li, C., Sun, P. Z., and Wu, Y. (2021). Preliminary demonstration of in vivo quasi-steady-state CEST postprocessing—correction of saturation time and relaxation delay for robust quantification of tumor MT and APT effects. *Magn. Reson. Med.* 86, 943–953. doi: 10.1002/mrm.28764

Zhao, X., Wen, Z., Huang, F., Lu, S., Wang, X., Hu, S., et al. (2011). Saturation power dependence of amide proton transfer image contrasts in human brain tumors and strokes at 3 T. *Magn. Reson. Med.* 66, 1033–1041. doi: 10.1002/mrm.22891

Zheng, J., Ramasinghe, S., and Lucey, S. (2021). Rethinking positional encoding. arXiv preprint arXiv:2107.02561.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/fnins.2023.1230814

OPEN ACCESS

EDITED BY
Shu Zhang,
Northwestern Polytechnical University, China

REVIEWED BY
Germain Arribarat,
INSERM U1214 Centre d'Imagerie Neuro Toulouse (ToNIC), France
Giovanni Mogicato,
Ecole Nationale Vétérinaire de Toulouse (ENVT), France
Xiao Li,
Northwest University, China
Timo Dickscheid,
Helmholtz Association of German Research Centres (HZ), Germany

*CORRESPONDENCE
Thierry Delzescaux
✉ thierry.delzescaux@cea.fr

RECEIVED 29 May 2023
ACCEPTED 31 October 2023
PUBLISHED 11 January 2024

CITATION
Piluso S, Souedet N, Jan C, Hérard A-S, Clouchoux C and Delzescaux T (2024) giRAff: an automated atlas segmentation tool adapted to single histological slices.
Front. Neurosci. 17:1230814.
doi: 10.3389/fnins.2023.1230814

giRAff: an automated atlas segmentation tool adapted to single histological slices

Sébastien Piluso[1,2], Nicolas Souedet[1], Caroline Jan[1], Anne-Sophie Hérard[1], Cédric Clouchoux[2] and Thierry Delzescaux[1]*

[1]Université Paris-Saclay, CEA, CNRS, MIRCen, Laboratoire des Maladies Neurodégénératives, Fontenay-aux-Roses, France, [2]WITSEE, Paris, France

Conventional histology of the brain remains the gold standard in the analysis of animal models. In most biological studies, standard protocols usually involve producing a limited number of histological slices to be analyzed. These slices are often selected into a specific anatomical region of interest or around a specific pathological lesion. Due to the lack of automated solutions to analyze such single slices, neurobiologists perform the segmentation of anatomical regions manually most of the time. Because the task is long, tedious, and operator-dependent, we propose an automated atlas segmentation method called giRAff, which combines rigid and affine registrations and is suitable for conventional histological protocols involving any number of single slices from a given mouse brain. In particular, the method has been tested on several routine experimental protocols involving different anatomical regions of different sizes and for several brains. For a given set of single slices, the method can automatically identify the corresponding slices in the mouse Allen atlas template with good accuracy and segmentations comparable to those of an expert. This versatile and generic method allows the segmentation of any single slice without additional anatomical context in about 1 min. Basically, our proposed giRAff method is an easy-to-use, rapid, and automated atlas segmentation tool compliant with a wide variety of standard histological protocols.

KEYWORDS

atlas segmentation, image registration, histology, brain, mouse

1 Introduction

In the last few decades, conventional histology has benefited from the expansion of light microscopy (Wilt et al., 2009; Ghaznavi et al., 2013; Milligan et al., 2019), in conjunction with the development of a wide range of biological staining techniques (Kuan et al., 2015; Kim et al., 2017; Erö et al., 2018; Tward et al., 2020; Wang et al., 2020). Cutting and acquisition protocols have become more and more sophisticated over time, providing a broad variety of procedures. This made it possible to observe the brain in an unprecedented way (Vandenberghe et al., 2016; Erö et al., 2018; Milligan et al., 2019; Tward et al., 2020). However, the resulting data remain massive and difficult to analyze for most of the labs. This is the case for the mouse brain in preclinical studies (Milligan et al., 2019).

Automated tools for analyzing these tissues, allowing the detection of biological objects and identification of the anatomical regions of interest (ROIs) to which they belong, are essential. Object segmentation has seen a tremendous upturn with the expansion of deep neural networks (Ronneberger et al., 2015; Falk et al., 2019). However, accurately identifying ROIs is still challenging and usually requires a brain atlas or expert knowledge of neuroanatomy.

As a result, many histological protocols are focused on specific anatomical regions, lesion areas, or pathological biomarkers, only on several well-chosen slices of interest within the brain (Lebenberg et al., 2010; Mesejo et al., 2012; Kim et al., 2015, 2017; Niedworok et al., 2016; Pagani et al., 2016; Renier et al., 2016; Ye et al., 2016; Dudeffant et al., 2017; Stolp et al., 2018; Zeng, 2018; Chen et al., 2019; Eastwood et al., 2019; Pallast et al., 2019; Bayraktar et al., 2020; Hérard et al., 2020; Sen et al., 2020; Song et al., 2020; Lam et al., 2022; Yee et al., 2022). It is prone to many drawbacks: this tedious work often yields non-reproducible operator-dependant results, suffers from inter- and intra-individual variability, and requires special attention in the statistical analysis design.

Digital mouse brain atlases aimed both to establish a rigorous, precise, and common reference of delineation for anatomical ROIs and, more importantly, to use them as a segmentation tool (Dauguet et al., 2007; Lein et al., 2007; Lau et al., 2008; Dubois et al., 2010; Johnson et al., 2010; Papp et al., 2014; Kuan et al., 2015; Tward et al., 2020; Wang et al., 2020).

A digital atlas being tree-dimensional (3D), the experimental volume needs to be reconstructed so that their respective dimensionality matches. But it is possible to reconstruct the organ in 3D using registration techniques when all or enough serial slices are cut and digitized (Ourselin et al., 2001; Modat et al., 2014; Agarwal et al., 2016; Niedworok et al., 2016; Fürth et al., 2018; Eastwood et al., 2019). This is the main issue to tackle, which cannot be achieved in most of the studies since protocols are not designed to yield 3D histology. One solution to overcome the lack of histological material is to use blockface photography (Toga et al., 1994) as a whole-brain template to achieve 3D reconstruction of several histological modalities of the same sample (Dauguet et al., 2007; Dubois et al., 2010; Vandenberghe et al., 2016). Indeed, 3D histology protocols are time-consuming, expensive, and neurobiologists often acquire only a limited number of slices. Therefore, the delineation of anatomical regions is mostly performed manually on the experimental data and/or the identification of their corresponding atlas slice is based on prior anatomical knowledge (Lebenberg et al., 2010; Ye et al., 2016; Iglesias et al., 2018; Pichat et al., 2018; Balakrishnan et al., 2019; Chen et al., 2019; Chon et al., 2019; Henderson et al., 2019; Pallast et al., 2019; Wu et al., 2019; Yates et al., 2019; Bayraktar et al., 2020; Hérard et al., 2020; Lam et al., 2022; Rodarie et al., 2022).

Furthermore, with the expansion of artificial intelligence techniques used to automatically segment brain slices, the need for reliable annotated database creation has dramatically increased in the last 5 years (de Vos et al., 2017, 2019; Krebs et al., 2017; Li and Fan, 2017; Rohé et al., 2017; Sokooti et al., 2017; Yang et al., 2017; Balakrishnan et al., 2018, 2019; Krepl et al., 2021; Sadeghi et al., 2022; Carey et al., 2023). Hence, automated, rapid, and adaptable atlas segmentation tools are still lacking but mandatory, for instance, when dealing with the segmentation of so-called *single* brain slices (devoid of 3D reference) needing to locate the 2D plane of each slice within a 3D atlas template volume. As the mouse brain has an elongated shape, most of the studies observe mouse brains in the coronal incidence (Bohland et al., 2010; Berlanga et al., 2011; Renier et al., 2016; Vandenberghe et al., 2016; Stæger et al., 2020), and we therefore focused on this incidence. Three parameters enable the exact location of a single slice plane within the atlas volume: (1) the *z-position* of the slice along the rostro-caudal (antero-posterior, AP) axis orthogonal to the coronal plane; (2) the tilting angle φ around the dorso-ventral (infero-superior, IS) axis; and (3) the tilting angle β around the transversal (left-right, LR) axis (Figure 1). Some tools, such as cutting matrices, can be used to obtain a quasi-perfect coronal cutting incidence, i.e., with φ and β tilting angles close to zero and therefore negligible, but usually, φ and β tilting angles lead to discrepancies when comparing "real life" slices and atlas ones.

Some studies focused on identifying possible tilting angles φ and β to refine 2D-plan location within the 3D atlas space (Xiong et al., 2018), while others proposed automated methods or user-friendly softwares to handle 2D slices within a 3D space toward *z-position*-oriented estimation (Puchades et al., 2019; Tappan et al., 2019). These strategies present a more or less accurate estimation of both tilting angles and are not fully automated since they all include manual processing to estimate the *z-position*. Basically, manual processing limits the use of such methods on a large scale for the study of mouse cohorts, in particular. More recently, a feature-based method called AMaSiNe was proposed to automatically estimate *z-position*, φ, and β (Song et al., 2020). Authors evaluate them with precision (< 100 µm), and segmentation results have been validated on two specific small regions only (primary visual area and dorsal lateral geniculate complex). However, the method is non-reproducible for the analysis of a single slice. In addition, the method is only robust from a minimum of three slices. Finally, a completely different approach has been proposed, using deep neural networks (Sadeghi et al., 2022; Carey et al., 2023). These methods require a large number of slices to train the network and rely on manual ground truth definition. Such estimates are prone to inter- and intra-individual error; their result is subjective and usually performed only on a relatively small part of the dataset. Moreover, the large variety of histological staining, along with the different imaging modalities, makes it very difficult to build up an exhaustive database to train a fairly generic neural network. Most of the existing methods are either very complex and not user-friendly (codes without interface) to be implemented by neurobiologists or require knowledge in neuroanatomy to be used appropriately, both greatly reducing their scope of application.

The method we propose is intended to be generic enough to be used by anyone and benefits from a user-friendly interface. The fully automatic mode we propose gives reliable results, and the user can still adjust parameters. We focused on the estimation of the *z-position* of single coronal slices. Our automated method is reproducible and can align and segment any number of single slices within a digital 3D atlas. Moreover, we developed a dedicated multi-slices extension to meet ROI-driven histological protocols, resulting in a set of slices from the same brain. Our method is based on a linear registration algorithm as well as an independent and multimodal similarity criterion. The Block Matching (BM) algorithm (Ourselin et al., 2001) was chosen as a robust strategy to register data from different modalities. This method was later included in the NiftyReg library (Modat et al., 2014) and is still well used in many applications (Niedworok et al., 2016; Iglesias et al., 2018; Balakrishnan et al., 2019; Borovec et al., 2020; Mancini et al., 2020). Normalized Mutual Information (NMI)

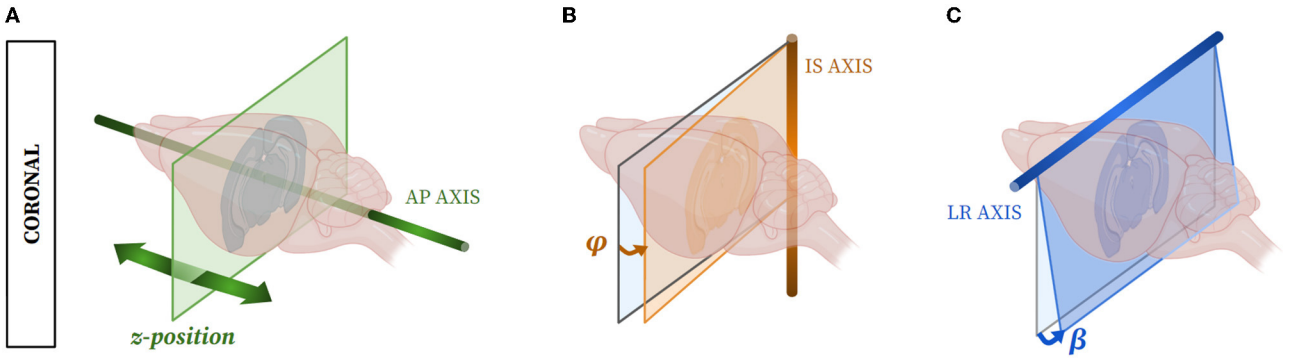


FIGURE 1
The three parameters used to exactly evaluate the 2D location of a coronal mouse brain slice within a template from an atlas: **(A)** the *z-position* along the AP axis, **(B)** the possible tilting angle φ around the IS axis, and **(C)** the possible tilting angle β around the LR axis.

(Studholme et al., 1998) was chosen as a robust similarity metric adapted to multimodality. This metric has proven its efficiency in many biomedical image processing applications (Jefferis et al., 2007; Geha et al., 2008; Dorocic et al., 2014; Costa et al., 2016). The idea of the method is to explore registrations between the experimental single slice and the ones from the atlas template, with increasing degrees of freedom. The NMI criterion is used to propose a generic evaluation framework of the relative similarity between slices after each step of registration. Basically, the method combines similarity information coming from *Rig*id and *Aff*ine registration, which explains the acronym we defined for this method: giRAff. We chose to refer to the Allen mouse Brain Atlas (ABA), a digital atlas widely used in neurobiology (Lein et al., 2007; Lau et al., 2008; Kuan et al., 2015). Also, we focused on histological slices covering the cortex, excluding the main olfactory bulb and the cerebellum. Most of the biological samples come from healthy subjects, but we also present some preliminary results on a pathological subject (Alzheimer's disease mouse model).

In addition, high-performance computing strategies were used to reach our goal of segmenting a large number of histological slices. Indeed, as registrations have a relatively high computational cost, calculations were distributed on hundreds of CPU cores through the dedicated tool SomaWorkflow (Laguitton et al., 2011). Finally, to make the method easy-to-use, it was implemented within the user-friendly open-source software interface BrainVISA (Cointepas et al., 2001; Lebenberg et al., 2010).

2 Materials and methods

2.1 Materials

2.1.1 Digital mouse brain atlas

In this study, we used the template and atlas from the Allen mouse Brain Atlas (ABA) (© 2015 Allen Institute for Brain Science. Allen Brain Atlas API. Available from: brain-map.org/api/index.html). It is composed of two perfectly aligned datasets: a template that represents the average anatomy of the mouse brain and labels that represent the theoretical delimitation of anatomical regions delineated by an expert on the template data. This template was built as an average autofluorescence of 1,675 serial two-photon tomography C57Bl/6J mouse brains, for which we considered each coronal slice $T_a \in B$ independently. B is the ensemble of slices describing the template volume considered a succession of independent slices in a given incidence (here coronal). The slice thickness is $e_t = 100$ µm and the in-plane resolution is 10×10 µm^2. In this study, we aimed to register 2D template images onto experimental histological slices. The purpose is to identify in the single slice of interest all the regions defined in the ABA reference corresponding slice.

2.1.2 Histological dataset

In this study, we aim to segment single 2D mouse brain coronal slices I_r, digitized from two different and independent histological modalities (see Supplementary material S0 for detailed protocols).

The first modality (so-called *autofluorescence*) is the autofluorescence of six clarified half mouse brains (M_1-M_6) imaged using light sheet fluorescence microscopy (Renier et al., 2016) that are considered as a succession of 2D coronal slices I_r devoid of cutting artifacts by nature. Those data were initially acquired with a resolution of $4 \times 4 \times 3$ µm^3 and resampled to $25 \times 25 \times 100$ µm^3 to generate a standard histological dataset.

The second modality (so-called *cresyl violet*) is cresyl violet-based Nissl staining of seven mouse whole brains produced in our laboratory (Vandenberghe et al., 2016) cut in the coronal incidence (I_r) using a microtome and digitized with a flatbed scanner. This second dataset includes six C57Bl/6J wild-type mouse brains (M_7-M_{12}) and one APP/PS1dE9 amyloid mouse brain (M_{13}), a transgenic mouse model of Alzheimer's disease (Dudeffant et al., 2017). The slice thickness is $e_r = 20$ µm (one every four slices) and the in-plane resolution is 25×25 µm^2. Regarding the cutting protocol, no specific instructions were given to prevent tilting angles. The cresyl violet data arose from our laboratory routine protocols in conventional histology.

2.2 Methods

2.2.1 Preprocessing

The template slices were first resampled in 2D to make the pixel size identical to the experimental data. Thus, the same number of pixels were used in the registration process by BM.

All images were resampled at $25 \times 25\ \mu m^2$ for registration. This resampling was chosen as a compromise between a pixel size small enough to apply the registration in a reasonable time and large enough to preserve sufficient details in the image for the registration algorithm. In such a conventional histology study, data are commonly resampled at an in-plane resolution of $25 \times 25\ \mu m^2$ (Renier et al., 2016) or $50 \times 50\ \mu m^2$ (Song et al., 2020).

The template slices were also manually centered to correspond to the experimental images. This gave a good initialization, minimizing the amplitude of the displacements induced by the registration process and maximizing the tissue overlap at an equivalent field of view.

2.2.2 The giRAff method for one single slice

The giRAff method estimates the *z-position* of a single mouse brain slice within an atlas volume at a given incidence and considers zero or negligible tilting angles. This estimated *z-position* is associated with a transformation resulting from the registration between the corresponding template slice at the *z-position* and the experimental slice. The estimation of the *z-position* is given by the optimum of a similarity criterion estimated between the experimental slice considered and a set of registered slices from the template. The final result is the atlas segmentation of the single experimental slice considered through the registered and identified corresponding label slice.

The method is based on the atlas from the ABA and the linear registration method by Block Matching (BM) based on the Crossed Correlation (CC) similarity metric with the default parameters given by Ourselin et al. (2001), designed for such a histological dataset. Normalized Mutual Information (NMI) is the independent metric that quantifies the similarity between the registered template slices and the experimental single slice considered in pairs.

Given an incidence (here coronal), consider I_r an experimental single slice to be segmented by atlas and T_a a slice from an ensemble B of slices describing the template volume considered as a succession of independent slices, such as $\{T_a \in B\}$. Let L_a be a slice from an ensemble A of slices describing the labels considered as a succession of independent slices, such as $\{L_a \in A\}$, A and B being in the same geometry and perfectly aligned. Let N be the number of considered template slices in a given incidence (along the AP, IS, or LR axis), $a \in \mathbb{N}^*$, going from 1 to N, the considered template slice number. Each template slice (from B) has its corresponding slice containing the labels (from A). Assume $z = \hat{a}$, the estimated position of the slice I_r within the template, i.e., the corresponding slice containing the labels. We chose to register template images (test) onto the experimental data (reference) to preserve the native geometry of the single slice (experimental) given as input by a user. Hence, labels will be mapped in the end onto the single slice to match its initial configuration.

The exploratory process for each image $T_a \in B$ is carried out in three steps (Figure 2), with RIG and AFF representing the rigid and affine transformation space, respectively:

(1) Rigid registration using BM (transformation $\hat{\theta}_{\text{RIG}}$) between I_r (reference) and T_a (test) from B, followed by an NMI similarity calculation S_{RIG} between the registered image $T_a \cdot \hat{\theta}_{\text{RIG}}$ and I_r,

$$\begin{aligned} S_{\text{RIG}}(I_r, T_a; \hat{\theta}_{\text{RIG}}) &= \text{NMI}(I_r, T_a \circ \hat{\theta}_{\text{RIG}}) \\ \text{with } \hat{\theta}_{\text{RIG}} &= \underset{\theta_{\text{RIG}} \in \text{RIG}}{argmax} \left(\text{CC}(I_r, T_a \circ \theta_{\text{RIG}})\right) \end{aligned} \quad (1)$$

(2) Affine registration using BM (transformation $\hat{\theta}_{\text{AFF}}$) between I_r (reference) and $T_a \cdot \hat{\theta}_{\text{RIG}}$ (test) registered in rigid (initialization), followed by NMI similarity calculation S_{AFF} between the registered image $T_a \cdot \hat{\theta}_{\text{RIG}} \cdot \hat{\theta}_{\text{AFF}}$ and I_r,

$$\begin{aligned} S_{\text{AFF}}(I_r, T_a; \hat{\theta}_{\text{AFF}}) &= \text{NMI}(I_r, T_a \circ \hat{\theta}_{\text{RIG}} \circ \hat{\theta}_{\text{AFF}}) \\ with\ \hat{\theta}_{\text{AFF}} &= \underset{\theta_{\text{AFF}} \in \text{AFF}}{argmax} \left(CC(I_r, T_a \circ \theta_{\text{RIG}} \circ \theta_{\text{AFF}})\right) \end{aligned} \quad (2)$$

(3) Calculation of the weighted average S_w from the two similarity values S_{RIG} and S_{AFF}:

$$\begin{aligned} S_w(I_r, T_a, \hat{\theta}_{\text{RIG}}, \hat{\theta}_{\text{AFF}}) &= (1-w)\, S_{\text{RIG}}(I_r, T_a; \hat{\theta}_{\text{RIG}}) \\ &\quad + w\, S_{\text{AFF}}\,(I_r, T_a; \hat{\theta}_{\text{RIG}}, \hat{\theta}_{\text{AFF}}) \end{aligned} \quad (3)$$

with $0 \leq w \leq 1$ the rigid-affine weighting.

From the weighted average S_w calculated for each slice T_a from B, a search of the maximum of similarity is performed to determine the slice number z from B, maximizing this similarity criterion from the N template slices:

$$z(I_r, B) = \underset{T_a \in B}{argmax}(S_w(I_r, T_a, \hat{\theta}_{\text{RIG}}, \hat{\theta}_{\text{AFF}})) \quad (4)$$

Thus, the result of the giRAff method can be summarized as follows:

$$\textbf{giRAff}(I_r, B) = (z, \widehat{\theta}_{\text{RIG}}, \widehat{\theta}_{\text{AFF}}) \quad (5)$$

The rigid and affine transformations $\hat{\theta}_{\text{RIG}}$ and $\hat{\theta}_{\text{AFF}}$ estimated by BM are successively applied at the slice $L_{\hat{a}}$ from the atlas at the position $\hat{a} = z$ to superimpose the registered image containing the labels $\hat{L}_{\hat{a}}$ on I_r, the experimental image.

$$\widehat{L}_{\hat{a}}(z, \widehat{\theta}_{\text{RIG}}, \widehat{\theta}_{\text{AFF}}) = L_z \circ \widehat{\theta}_{\text{RIG}} \circ \widehat{\theta}_{\text{AFF}} \quad (6)$$

The transformation matrices $\hat{\theta}_{\text{RIG}}$ and $\hat{\theta}_{\text{AFF}}$ are applied to the slice $L_{\hat{a}}$ with the nearest neighbor interpolation to preserve the initial values of the labels. The experimental single slice I_r is then automatically segmented by the ABA. Quantitative region-based analysis can then be carried out on it thanks to the method.

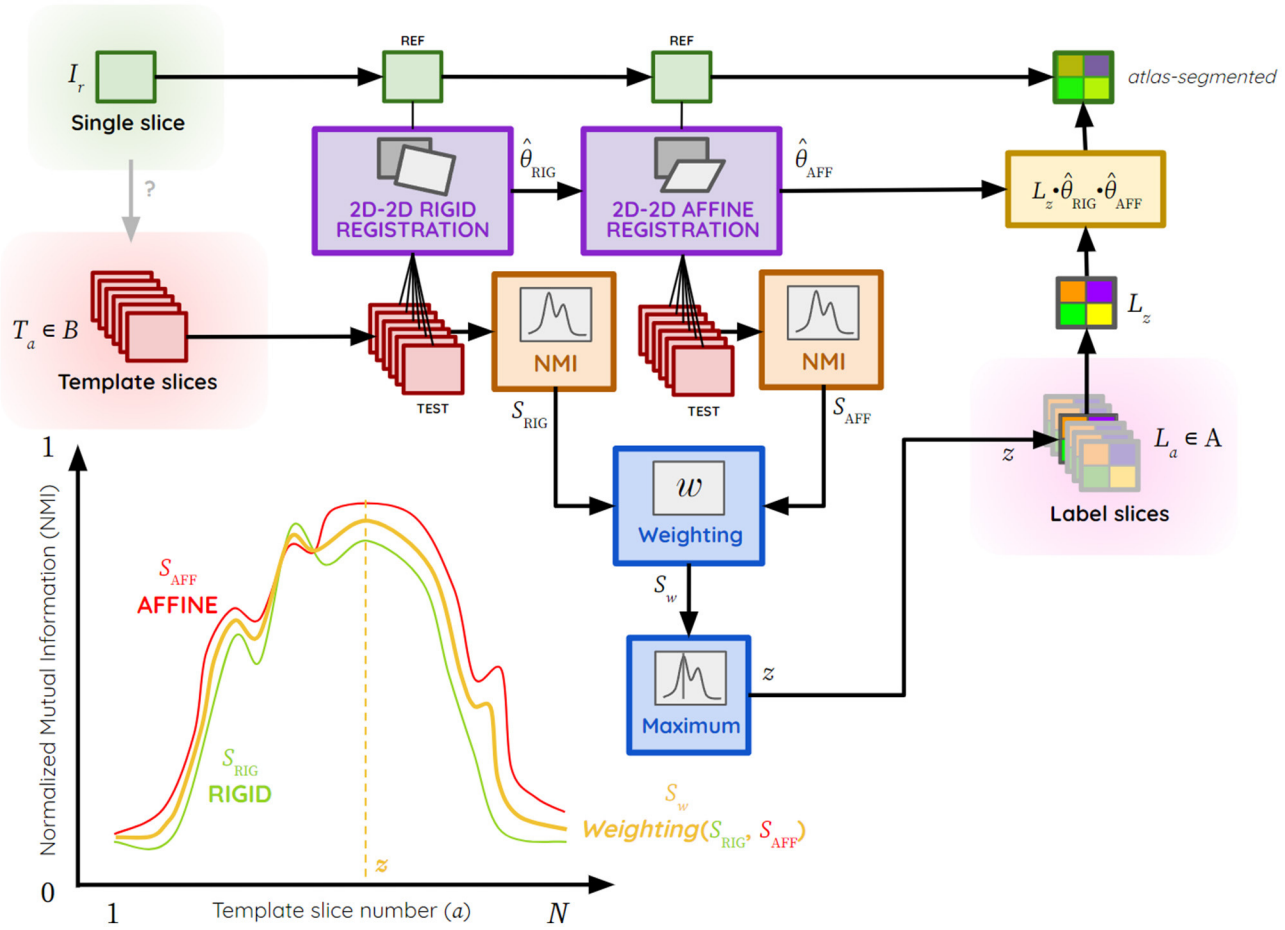


FIGURE 2
Atlas segmentation process of a single brain slice I_r by the giRAff method using template slices $T_a \in B$, theoretical example of NMI curve plots with S_{RIG} (green), S_{AFF} (red) and S_w (yellow), z the template slice number corresponding to a slice I_r estimated by the giRAff method, $\hat{\theta}_{RIG}$ the rigid transformation, $\hat{\theta}_{AFF}$ the affine transformation, B the ensemble of template slices, and $L_a \in A$ the ensemble of label slices matching the template ones.

2.2.3 The giRAff$_m$ extension for a multi-slices case

2.2.3.1 Relative scaling factor between brain samples

Two mouse brains are often considered to be roughly the same size, but this is not the case in practice. Two factors influence the size of the organ, in particular: inter-individual variability (natural) and the extraction, cutting, and staining protocol to which the sample is subjected before analysis (non-natural).

Let us consider a multi-slices case, i.e., a series of single histological slices from the same mouse brain not enabling its 3D reconstruction. Let d_r be the constant inter-slice distance between single slices from the experimental volume. Let d_t be the inter-slice distance between slices from the template volume. To realistically estimate the corresponding distance d_r depending on d_t in the template, the differences in brain volumetrics must be taken into account. Not taking them into account would lead to a deviation in the estimation of successive slice positions (Figure 3A). For this reason, a *relative scaling factor* γ (RSF) was introduced, which reflects the size difference between an experimental brain and the atlas template volume on the axis to which the considered incidence plane is orthogonal (Figure 3B). The affine registration automatically corrects the scaling factors in the other two directions (α, β) relative to the plane of incidence considered. This RSF γ is relative because no modality accounts for an absolute reference geometry: it is relative between two modalities.

Thus, assume:

$$\begin{cases} 0 < \gamma < 1 & \Leftrightarrow \text{ shrinkage} \\ \gamma = 1 & \Leftrightarrow \text{ same size} \\ 1 < \gamma < +\infty & \Leftrightarrow \text{ enlargment} \end{cases}$$

The distances d_r and d_t are defined as a function of γ and the two slice thicknesses e_r for the experimental data and e_t for the template data:

$$d_r = \gamma \frac{e_r}{e_t} d_t \tag{7}$$

Let $r \in N^*$ be the slice number from the experimental data ranging from $r = 1$ to $r = M$, and t the slice number estimated to

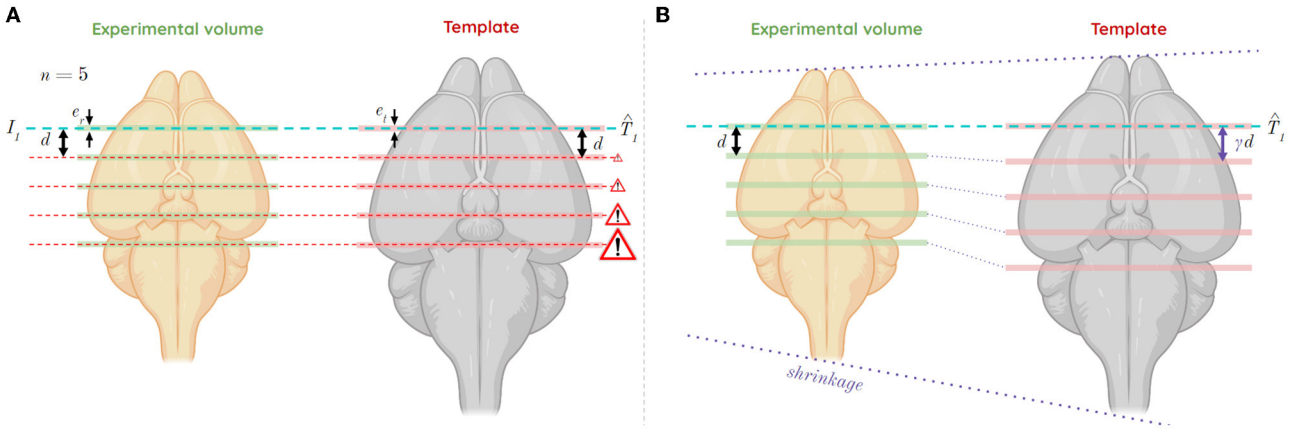


FIGURE 3
Shrinkage case of the experimental volume compared to the template volume, and its impact in a multi-slices study in the coronal incidence along the AP axis, for slice thicknesses e_r and e_t from the experimental and template volumes, respectively, as well as an inter-slice distance d between each considered slice from the experimental volume. **(A)** Study of five slices from a first T_1 estimate without consideration of shrinkage between volumes (mismatch). **(B)** Study of five slices from a first estimate T_1 considering the shrinkage between volumes using RSF γ (correction).

be the most similar in the template by the giRAff method, ranging from $t = 1$ to $t = N$. The equation of the affine line linking slice numbers from the two volumes to each other can thus be deduced:

$$\hat{t}(r) = \gamma \frac{e_t}{e_r}(r-1) + \hat{t}_1 \tag{8}$$

with $\hat{t}_1$ the y-intercept corresponding to the result of the giRAff method applied to the first slice of the experimental volume studied ($r = 1$).

2.2.3.2 Operating mode

For each considered experimental single slice from a multi-slices set, similarity values with all the template slices are computed by the giRAff method and stored in a list s_w (see Equation 3, which is applied for each slice $T_a \in$ B). The multi-slices analysis aims to bring each of these lists into a single referential to pool their contribution.

Assume $(u_s)_{s\in\mathrm{N}^*}$ an arithmetic series determining the first template slice number to be tested in the case of a multi-slices study and $(v_s)_{s\in\mathrm{N}^*}$ an arithmetic series determining the last template slice number to be tested in the case of a multi-slices study, we then have:

$$u_s = u_1 + \frac{d_t}{e_t}(s-1) \; and \; v_s = N - \frac{d_t}{e_t}(n-s) \qquad (9) \; and \; (10)$$

with $u_1 = 1$ corresponding to the first template slice number.

Values from the series $(u_s)_{s\in\mathrm{N}^*}$ and $(v_s)_{s\in\mathrm{N}^*}$ are rounded to the nearest integer so that they correspond to real slice numbers.

The giRAff method is successively executed for each slice s, solely on a range of template slices $B_{[us;vs]} \subset B$ defined by the two series. This range contains the same number of slices d_t/e_t rounded off to the nearest unit. This amounts to determining the *z-position* of the first studied slice from the mutualization of the similarity information S_w of all the slices in the multi-slices set. Once this *z-position* has been estimated in a common manner, it is propagated to the other slices of the series to determine their respective *z-positions*. The position of the other slices is deducted by adding the distance d_t in the template corresponding to the distance d_r, which separates the slices from each other in the experimental volume. Assuming z_m is the *z-position* estimated by combining different similarity information in the multi-slices case, as with the classical giRAff method, a calculation of the maximum similarity is then performed to determine the desired position z_m:

$$z_m(E,B) = \underset{T_a \in B_{[u_s;v_s]}}{argmax} \left(\frac{1}{n} \sum_{s=1}^{n} \pi_s S_w(I_s, B_{[u_s;v_s]}, \hat{\theta}_{\mathrm{RIGs}}, \hat{\theta}_{\mathrm{AFFs}}) \right) \tag{11}$$

with S_w being a list containing the averaged NMI values for rigid and affine registration (see Equation 3), E a multi-slices ensemble, and π_s the contribution rate for each slice s ($\pi_s = 1/n$ by default, giving an equal contribution for each slice).

Assume giRAff$_\mathrm{m}$ is the extension of the giRAff method to a multi-slices study, which is defined as:

$$\mathbf{giRAff_m}(E,B) = (z_s, \widehat{\theta}_{\mathrm{RIGs}}, \widehat{\theta}_{\mathrm{AFFs}}) \tag{12}$$

For each slice I_s from E, a z_s position (deduced from z_m) as well as rigid and affine transformations $\hat{\theta}_{\mathrm{RIGs}}$ and $\hat{\theta}_{\mathrm{AFFs}}$ are determined, which allows the identified label slice $L_{\hat{a}}$ to be mapped onto the experimental single slice I_s. The contribution of each slice I_s can be adjusted toward the weight π_s. For example, if a slice I_s has many artifacts that might compromise the registration with the template slices, it is possible to manually adjust its influence by decreasing its contribution π_s or even remove it from the z_m estimation ($\pi_s = 0$).

The numbers $z_s = t_s$ of each of the slices from the multi-slices study from E can directly be calculated from z_m:

$$z_s = z_m + s\, d_t = \widehat{t_1} + s\frac{e_t}{\gamma e_r} d_r \tag{13}$$

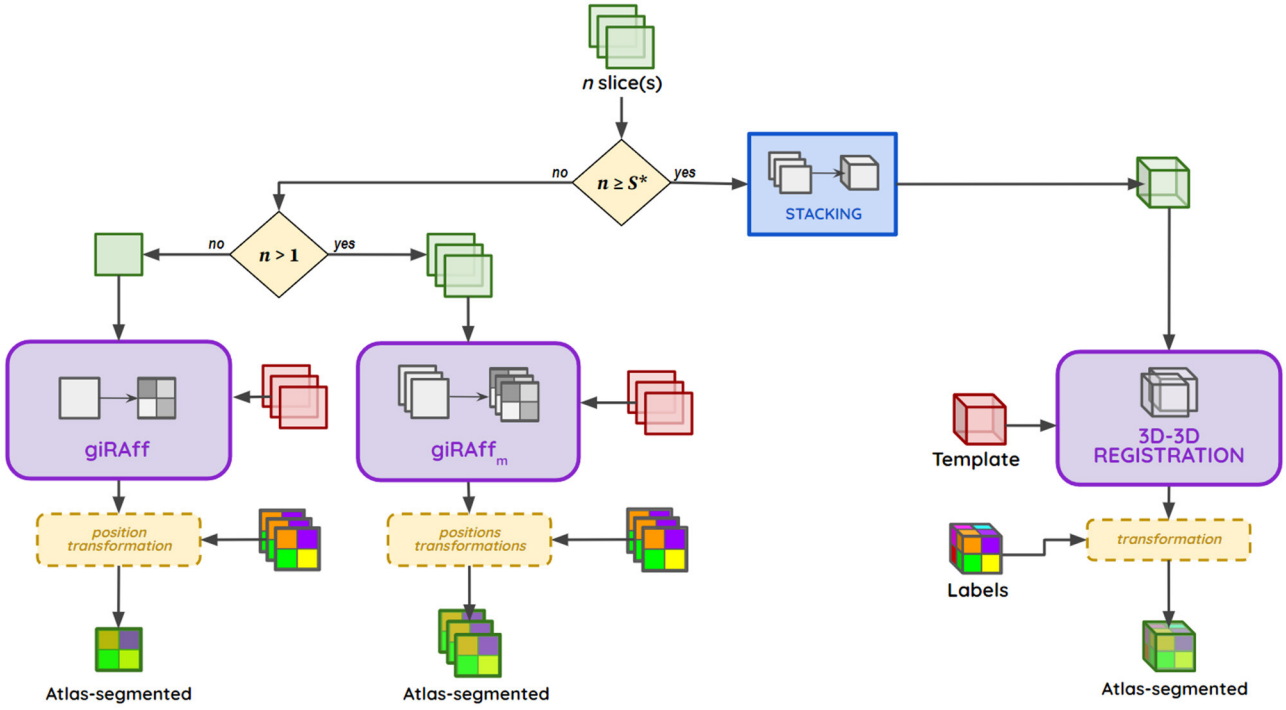


FIGURE 4
giRAffMapper generic pipeline performing the automated atlas segmentation of any number *n* of histological slices using the giRAff method as well as its extension giRAff$_m$. *S** represents the needed and required number of slices to reconstruct a 3D brain.

with $\hat{f}_s$ rounded to correspond to real slice numbers (integers). An affine transformation $\hat{\theta}_{\mathrm{AFF}s}$ is associated with each position z_s.

2.2.4 giRAffMapper: a generic pipeline

The generic pipeline giRAffMapper automatically performs the atlas segmentation of any number of slices corresponding to any histological experimental protocol. Let S^* be the needed and required number of slices to reconstruct a 3D brain. Whether it is for the analysis of a single slice ($n = 1$), for several slices in the analysis of a particular anatomical region ($1 < n < S^*$), or for a large enough number of slices to perform a 3D reconstruction of the brain ($n \geq S^*$), the giRAffMapper generic pipeline automatically processes any histological brain slice protocols (Figure 4).

2.2.5 Validation of the method

2.2.5.1 Metrics of validation

As our aim is to achieve an atlas segmentation as accurate as that of experts, we took the quantitative results of a neuroanatomist's evaluation as a reference. We asked an expert to identify the right number (*z-position*) of the template slice being the most similar to each experimental considered slice I_r, the so-called *Expert Rating* for the *z-position* (ER_z) (see Supplementary material S1). This made it possible to define the deviation of the *z-position* Δ_{sn} between ER_z and the *z-position* estimated by giRAff:

$$\Delta_{sn}(I_r, B) = \left| ER_z(I_r, B) - z(I_r, B) \right| \quad (14)$$

The final purpose being the segmentation of anatomical regions, we also calculated dice scores (Dice, 1945) between the manual segmentation from an expert on the experimental considered slice and the one resulting from the identified and registered template slice by the giRAff method. We then compared these obtained dice scores to those calculated after prior identification of the *z-position* by an expert.

2.2.5.2 Realistic histological protocols to perform region-based analysis

We designed realistic region-based histological protocols from mouse whole brain histological datasets with an expert. Six main regions of interest were chosen from different sizes and locations in the brain: cortex, striatum, hippocampus, thalamus, globus pallidus, and substantia nigra. We especially selected them because of their known involvement in neurodegeneration, especially concerning Alzheimer's, Parkinson's, or Huntington's diseases (Dostrovsky et al., 2002; Picconi et al., 2005; Teichmann et al., 2005). For each anatomical region, the protocol includes the identification of the respective slices in which this region starts and ends along the AP axis, as well as the number of slices to be considered and their inter-slice distance, allowing quantitative studies (see Supplementary material S2). To assess the robustness of such an exploratory approach, we tested all possible protocol combinations covering each region and brain considered, given a constant inter-slice distance.

2.2.6 Determination of the rigid-affine weighting *w* for a given imaging modality

To determine the optimal rigid-affine weighting to be applied for a given imaging modality, we evaluated the average Δ_{sn} values for each possible weighting, using steps of 0.01, for all the slices from the brains in a given modality. From this evaluation, we estimated an average curve of Δ_{sn} as a function of w, which gave us an average trend displaying which rigid-affine weighting w minimizes deviation Δ_{sn} and, therefore, maximizes the accuracy of the method. To get a realistic idea of this trend for conventional histological slices, it is necessary to exclude from the overall estimate brains suffering from too many artifacts (air bubbles, tearing, missing tissue, etc.) that could compromise this evaluation.

2.3 Implementation details and source code

Considering the large number of calculations, the pipeline was run using distributed computing on multiple microprocessors using the SomaWorkflow library of BrainVISA software (Laguitton et al., 2011). BrainVISA is an open-source software platform for neuroimaging research, including visualization tools and graphical user interfaces (https://brainvisa.info). This study was conducted on a workstation Ubuntu 16.04; LTS 64-bits; Intel® Xeon® CPU E5-2620 v2 @ 2.10GHz × 24 (24 computing cores); 128 GB of Random Access Memory (RAM), with the support of our Titan2 calculator composed of five DELL R610 bi-processor nodes on Intel® Xeon® CPU X5675 @ 3.07GHz × 12 and 48 Go of RAM, one DELL R610 bi-processor node on Intel® Xeon® CPU X5667 @ 3.07GHz × 8 and 48 Go of RAM, and six DELL R630 bi-processor nodes on Intel® Xeon® CPU E5-2630 v3 @ 2.40GHz × 16 and 128 Go of RAM (representing 328 computing cores overall).

3 Results

The giRAff method has the advantage of being exhaustive in exploring all the possible correspondences after linear registration between a single slice under study and the slices from the average template. This exploration is performed in a minimum of time thanks to a distributed implementation. The choice of the registration algorithm as well as the similarity metric was made to suit multimodal studies, and their independence provides robustness in the identification of the right *z-position* for a given single slice.

The giRAff$_m$ extension has been specially designed for multi-slices studies, where the RSF is taken into account for an accurate and realistic estimation of the common *z-position* for a given dataset.

All these developments are gathered in a generic pipeline able to automatically segment any number of slices by atlas. The method presents the advantage of being embedded in an easy-to-use software for simple utilization (see Supplementary material S6).

We used two complementary metrics to evaluate the efficiency of the method in its two different aims: its ability to identify the right *z-position* of single histological slices, whatever their number, and its ability to present relatively good atlas segmentation scores after registration.

3.1 Single histological slice segmentation by giRAff

3.1.1 Determination of the rigid-affine weighting *w*

We first evaluated which rigid-affine weight w minimizes the Δ_{sn} criterion for each modality: the autofluorescence (Figure 5A) and the cresyl violet (Figure 6A). For the autofluorescence, the trend was clearly not toward a rigid-affine weighting w at extremes (0 or 1). No particular weighting appeared to be especially optimal between these extreme values. We therefore chose a rigid-affine weighting $w = 0.50$ for this modality to ensure robustness in the use of the two types of registration and to avoid the extreme weightings, which can be a source of misidentification (high Δ_{sn}). Concerning cresyl violet, it was necessary to remove data presenting too important artifacts (M_7), making them non-representative for the evaluation of the global trend of the rigid-affine weighting w. In contrast to autofluorescence modality, a clear trend appeared in favor of a weighting $w = 1$ for the cresyl violet, which minimized mean Δ_{sn}. This means that the NMI resulting from affine registration prevailed for this imaging modality in the estimation of the *z-position* of single slices in comparison to an expert.

3.1.2 Precision and robustness of the method

The giRAff method was applied independently on 2,135 single half-slices (one hemisphere) and 636 whole slices (whole brain) from two modalities from 13 mouse brains. In routine protocols performed in our laboratory, φ and β angles were estimated below 5° (see Supplementary material S3) and were neglected in this study. The deviation Δ_{sn} compared to an expert was calculated for every single slice considered from this dataset. The giRAff method was able to identify any single mouse brain slice with an average accuracy of 1.20 ± 1.19 and 2.05 ± 3.05 slices for the autofluorescence and the cresyl violet, respectively (Table 1). This represented an average precision of the *z-position* identification between 120 and 164 µm, respectively.

Concerning the autofluorescence, no high Δ_{sn} scores appeared, being mainly narrow around 0 and 200 µm, the largest deviation of 10 slices being obtained only once (M_1) among the six brains (Figure 5B). If we look qualitatively at the segmentation of the anatomical regions of interest, we notice that their delineation is close to that performed by an expert on the experimental slice (Figure 5C). From the smallest of the regions studied (substantia nigra) to the most elongated (cortex), the segmented shapes were quite close. These results were confirmed quantitatively by the dice scores (Figure 5D; see Supplementary material S4) evaluated on five slices among the brain M_1, which demonstrated the capacity of the giRAff method to obtain fairly high scores (around 0.90) after identification of the *z-position* for a given experimental slice. More importantly, those dice scores were widely comparable to those of an expert.

10.3389/fnins.2023.1230814

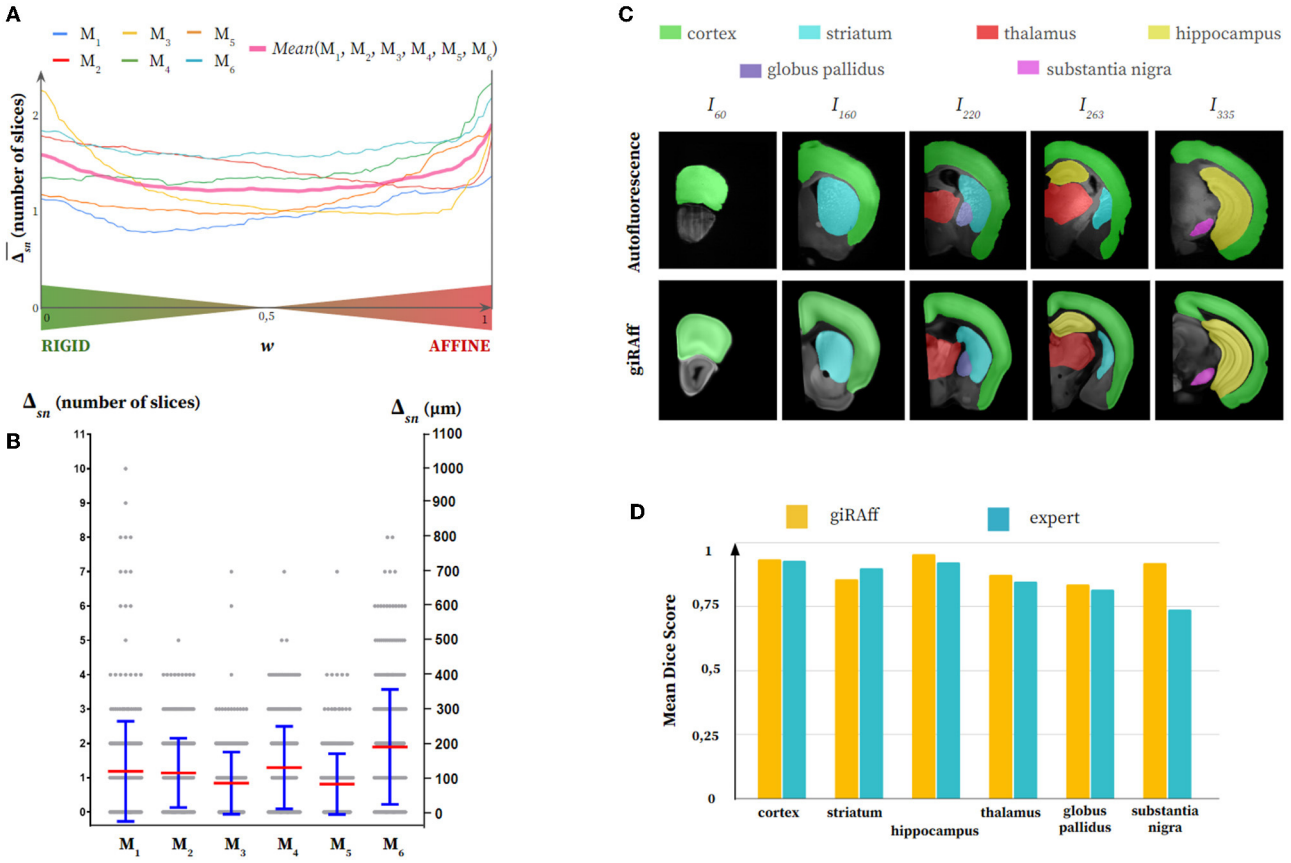


FIGURE 5
Single slice manual segmentation of the Autofluorescence half mouse brains (M_1-M_6): **(A)** Averaged Δ_{sn} values after application of the giRAff method for each rigid-affine weighting *w* from 0 to 1 by 1% increments, **(B)** Δ_{sn} values (gray) after application of the giRAff method (mean in red and standard deviation in blue) for each single slice, **(C)** segmentation of six anatomical regions of interest of various sizes (cortex, striatum, hippocampus, thalamus, globus pallidus, and substantia nigra) by an expert in five experimental single slices across the brain M_1 (first row) as well as their corresponding registered template slice identified by the giRAff method (second row), and **(D)** non-weighted mean dice scores (see Supplementary material S4) evaluated on the six anatomical regions for the five slices identified and registered from **(C)** between the giRAff method (yellow) and an expert (cyan).

Results for cresyl violet appeared to be somewhat less accurate, with Δ_{sn} scores being concentrated more between 0 and 300 µm on average, with the exception of M_7, for which values were significantly higher (Figure 6B). Misidentifications above 10 slices of deviation were also rare. The qualitative analysis of the segmentations showed that the anatomical regions corresponded rather well, with some small differences, in proportion for the substantia nigra or in shape for the striatum (Figure 6C). These small differences had very little impact on the dice score, which remained globally quite high (around 0.85), except for the substantia nigra and the globus pallidus (around 0.75). Similarly and most importantly, dice scores showed that segmentation results using the giRAff method on these five slices were still widely comparable to those of an expert (Figure 6D; see Supplementary material S4).

All in all, no particular difference in the accuracy of the giRAff method was noticed in identifying the *z-position* of slices from a brain including pathological lesions (M_{13}) compared to other brains (M_7-M_{12}): average Δ_{sn} and standard deviation (1.13 ± 0.91 slices) of M_{13} were significantly inferior to the mean evaluation on the whole cresyl violet dataset (2.05 ± 3.05 slices).

3.2 Multi-slices histological segmentation by giRAff$_m$

Based on the rigid-affine weighting empirically determined for each of the two modalities, the giRAff$_m$ extension was applied on multi-slices datasets based on routine histological sectioning protocols. Those protocols were designed by experts to correspond to studies of particular anatomical regions of different sizes in the coronal incidence: cortex, striatum, thalamus, hippocampus, globus pallidus, and substantia nigra (see Supplementary material S2). These conventional protocols involved a number of slices and an inter-slice distance, with the first slice of a given region being shifted at each iteration so that the entirety of the

10.3389/fnins.2023.1230814

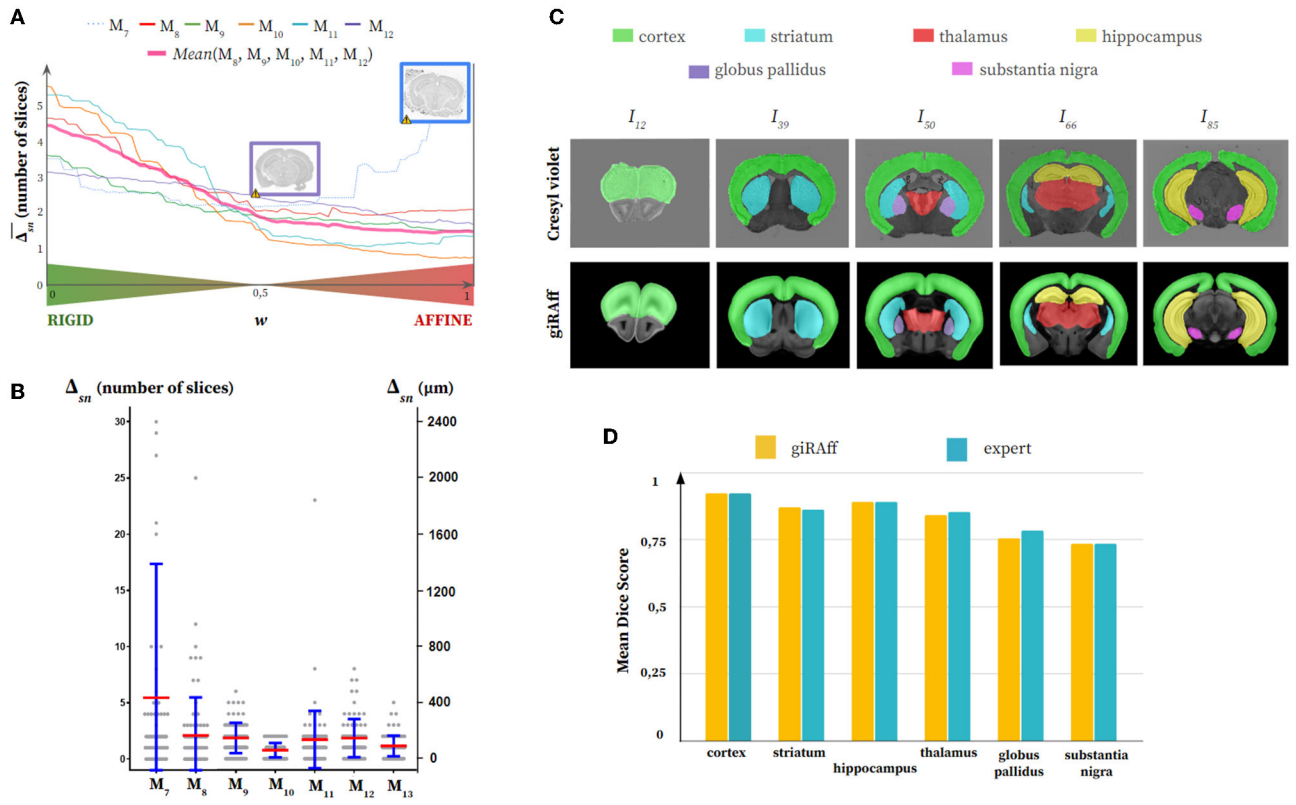


FIGURE 6
Single slice manual segmentation of the seven cresyl violet mouse brains (M_7-M_{13}): **(A)** Averaged Δ_{sn} values after application of the giRAff method for each rigid-affine weighting w from 0 to 1 by 1% increments, **(B)** Δ_{sn} values (gray) after application of the giRAff method (mean in red and standard deviation in blue) for each single slice, **(C)** segmentation of six anatomical regions of interest of various sizes (cortex, striatum, hippocampus, thalamus, globus pallidus, and substantia nigra) by an expert in five experimental single slices across the brain M_{10} as well as their corresponding registered template slice identified by the giRAff method, and **(D)** non-weighted mean dice scores (see Supplementary material S4) evaluated on the six anatomical regions for the five slices from **(C)** between the giRAff method (yellow) and an expert (cyan).

slices constituting each region were tested. The deviation Δ_{sn} was calculated for each slice included in every multi-slices case, and the result was averaged per anatomical region studied.

The deviation Δ_{sn} was estimated in three different contexts: (1) using the giRAff method considering each slice as single (same case as in Section 3.1 focused on the slices including each anatomical region considered), (2) using the giRAff$_m$ extension considering multi-slices protocols and an RSF γ_{Mi} evaluated for each brain thanks to the ER_z, and (3) using the giRAff$_m$ extension considering multi-slices protocols and an averaged RSF γ_m evaluated for a given protocol and imaging modality (see Supplementary material S5).

Concerning the autofluorescence, first, the multi-slices approach significantly reduced the average deviation Δ_{sn} and its dispersion, in general: Δ_{sn} criterion underwent a reduction between 55 and 105 µm and the standard deviation between 53 and 87% depending on the region, on average (Table 2; Figure 7A). The case of considering the RSF γ_{Mi} specific to each volume M_i (i ranging from 1 to 6) presented smaller deviations Δ_{sn} than the case of an average RSF γ_m (increase of the order of 8%). Depending on the experimental conditions that were applied to each volume, considering γ_{Mi} specific to each of them made it possible to obtain better accuracy in the detection of the z_m position. Estimating an accurate value of this RSF γ increased the precision of detecting the right position z_m by the giRAff$_m$ extension. On average, over all regions, the accuracy of z_m position detection in the multi-slices case by the giRAff$_m$ extension was equal to 57 ± 49 µm with γ_{Mi} and 63 ± 52 µm with γ_m for the autofluorescence.

Regarding the cresyl violet, second, the multi-slices approach strongly decreased the average deviation Δ_{sn} and its dispersion in general: Δ_{sn} criterion underwent a reduction between 53 and 169 µm, and the standard deviation between 69 and 90% depending on the region, on average (Table 2; Figure 7B). The use of the giRAff$_m$ extension in the multi-slices case significantly improved the overall detection accuracy of the z_m position in this modality. In contrast to what was observed for the autofluorescence data, the γ_m case presented better results (Δ_{sn} decreased by 4% on average over all regions) than for the consideration of the respective γ_{Mi}. Only the cortex region showed $\Delta_{sn}(\gamma_{Mi}) > \Delta_{sn}(\gamma_m)$ by 7%. For the other regions, considering γ_m rather than γ_{Mi} improved the detection of the correct *z-position* by 4% (striatum) to 27% (globus pallidus). On average, over all regions, the accuracy of z_m position detection in the multi-slices case by the giRAff$_m$ method was 94 ± 54 µm with γ_m. Whatever the case considered, the substantia nigra was the only region with high deviations: the accuracy Δ_{sn} was

TABLE 1 Single slices—autofluorescence and cresyl violet.

Autofluorescence		M_1	M_2	M_3	M_4	M_5	M_6	MEAN	
$\overline{\Delta_{sn}}$(nb of slices and µm)	µ	1.19	1.14	0.84	1.30	0.82	1.90	1.20	
		119 µm	114 µm	84 µm	130 µm	82 µm	190 µm	120 µm	
	σ	± 1.46	± 1.01	± 0.91	± 1.20	± 0.89	± 1.67	± 1.19	
		± 146 µm	± 101 µm	± 91 µm	± 120 µm	± 89 µm	± 167 µm	± 119 µm	
M (nb of slices per brain)		354	379	341	342	362	357	2,135	
Cresyl violet		M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}	MEAN
$\overline{\Delta_{sn}}$(nb of slices and µm)	µ	5.40	2.07	1.85	0.76	1.70	1.82	1.13	2.05
		432 µm	166 µm	148 µm	61 µm	136 µm	146 µm	90 µm	164 µm
	σ	± 11.84	± 3.38	± 1.35	± 0.64	± 2.54	± 1.69	± 0.91	± 3.05
		± 947 µm	± 270 µm	± 108 µm	± 51 µm	± 203 µm	± 135 µm	± 73 µm	± 244 µm
M (nb of slices per brain)		82	93	95	97	93	85	91	636

Average Δ_{sn} scores and standard deviation resulting from the application of the giRAff method compared to expert *z-position* evaluation (ER_z), as well as the number of slices M per brain on which it has been estimated for the six autofluorescence half mouse brains (M_1-M_6) and the seven cresyl violet mouse brains (M_7-M_{13}). The MEAN column presents the non-weighted average Δ_{sn} values for all the considered brains.

180 ± 40 µm while it was always < 80 µm for all other regions. Atlas segmentation of small anatomical regions was more challenging than for large regions, both for experts and for the proposed method.

A gain in accuracy was clearly observed when using the giRAff$_m$ extension compared to the giRAff method for the same slices considered independently: with a few exceptions, Δ_{sn} was brought down between 0 and 100 µm on average, whatever the modality and the region.

For one single slice, the giRAff method proposed an automated atlas segmentation in about 1 min using Titan2 infrastructure.

3.3 Cross-talk between giRAff and giRAff$_m$

Several single slices from cresyl violet mouse brains (M_7-M_{13}) suffered from histological artifacts. In most cases, the presence of a considerable artifact prevents segmentation of the entire histological slice. Such a slice is often discarded, or its segmentation is carried out manually if the damaged part does not concern the tissue of interest. Despite some considerable artifacts, the giRAff$_m$ extension still allows for identification of the correct *z-position* and segment the rest of the slice correctly. Some examples including such artifacts (tissue folding, missing tissue, and external noise) are presented in Figure 8.

4 Discussion

In this study, we proposed a method to automatically segment one or a set of single slices using a 3D digital atlas. The giRAff method, based on linear registration tools and on the NMI as a similarity metric, showed its ability to deal with any number of slices, adapting to very different standard histological protocols (3D fluorescence and 2D brightfield imaging). We demonstrated the robustness and the efficiency of the method by applying it on two different datasets: autofluorescence data, which was not affected by cutting artifacts, and histological slices from routine experimental protocols. It was indeed able to identify, depending on the protocol considered, the *z-position* of one or more single slice(s) with an accuracy of the order of one slice within the atlas template. This amounted to an identification deviation of less than about 100 µm on average, with dice scores comparable to those obtained by an expert. The method also showed its ability to deal with slices suffering from histological artifacts using the multi-slices approach.

The method was based on a balanced use of the similarity information evaluated after rigid and affine registration in an exploratory approach. In this context, the rigid-affine weighting w was of crucial importance as it allowed to adjust the use of NMI information to take advantage of the benefits from each type of registration. Indeed, in the exploratory approach we proposed, the two types of registration can be complementary. Rigid registration is often rough and avoids the identification of a particular slice that is the closest to the single slice considered, whereas affine registration makes the difference in improving tissue registration thanks to a greater number of degrees of freedom (shearing and scaling). On the contrary, affine registration could make slices correspond to each other with an inappropriate superposition of tissues forced by large deformations, whereas rigid registration does not allow such modifications, limits the deformations, and permits the differentiation of these slices. The use of a weighted proportion of the similarity information created a robust study framework for their comparison in an exploratory context. This represents a useful parameter to tune according to the amplitudes of the deformations considered or according to the biological protocol used. For the two modalities tested in this study, the trend was toward either 0.5 or 1. What we would suggest for users is to consider one or the other of the rigid-affine weighting given in the manuscript by default for their own data according to the imaging modality chosen. In the case of another specific imaging modality or for any doubt on the rigid-affine weighing chosen, the operator could easily test adjusting it from 0.5 to 1 or from 1 to 0.5. If

TABLE 2 Multi-slices—autofluorescence and cresyl violet.

Autofluorescence Δ_{sn} (nb of slices and µm)		Cortex	Striatum	Thalamus	Hippocampus	Globus pallidus	Substantia nigra	MEAN
giRAff	µ	1.59	1.28	1.14	1.16	1.11	1.33	**1.27**
		159 µm	128 µm	114 µm	116 µm	111 µm	133 µm	**127 µm**
	σ	± 3.68	± 1.30	± 1.12	± 1.03	± 1.15	± 1.06	**± 1.56**
		± 368 µm	± 130 µm	± 112 µm	± 103 µm	± 115 µm	± 106 µm	**± 156 µm**
giRAff$_m$ | γ_{Mi}	µ	0.54	0.63	0.52	0.57	0.53	0.63	**0.57**
		54 µm	63 µm	52 µm	57 µm	53 µm	63 µm	**57 µm**
	σ	± 0.49	± 0.48	± 0.47	± 0.48	± 0.52	± 0.50	**± 0.49**
		± 49 µm	± 48 µm	± 47 µm	± 48 µm	± 52 µm	± 50 µm	**± 49 µm**
giRAff$_m$ | γ_m	µ	0.72	0.73	0.57	0.60	0.51	0.66	**0.63**
		72 µm	73 µm	57 µm	60 µm	51 µm	66 µm	**63 µm**
	σ	± 0.62	± 0.54	± 0.48	± 0.48	± 0.51	± 0.50	**± 0.52**
		± 62 µm	± 54 µm	± 48 µm	± 48 µm	± 51 µm	± 50 µm	**± 52 µm**
Cresyl violet Δ_{sn} (nb of slices and µm)		Cortex	Striatum	Thalamus	Hippocampus	Globus pallidus	Substantia nigra	MEAN
giRAff	µ	2.22	1.54	2.62	2.79	1.55	4.38	**2.31**
		178 µm	123 µm	210 µm	223 µm	124 µm	350 µm	**286 µm**
	σ	± 3.41	± 1.90	± 3.65	± 3.84	± 1.55	± 4.99	**± 3.17**
		± 273 µm	± 152 µm	± 292 µm	± 307 µm	± 124 µm	± 399 µm	**± 254 µm**
giRAff$_m$ | γ_{Mi}	µ	0.84	0.84	1.05	0.96	0.89	2.45	**0.98**
		67 µm	67 µm	84 µm	77 µm	71 µm	196 µm	**78 µm**
	σ	± 0.51	± 0.46	± 0.44	± 0.50	± 0.48	± 0.57	**± 0.49**
		± 41 µm	± 37 µm	± 35 µm	± 40 µm	± 38 µm	± 46 µm	**± 39 µm**
giRAff$_m$ | γ_m	µ	0.90	0.81	0.94	0.87	0.65	2.27	**0.94**
		72 µm	65 µm	75 µm	70 µm	52 µm	182 µm	**75 µm**
	σ	± 0.60	± 0.53	± 0.51	± 0.51	± 0.47	± 0.50	**± 0.54**
		± 48 µm	± 42 µm	± 41 µm	± 41 µm	± 38 µm	± 40 µm	**± 43 µm**

Non-weighted average Δ_{sn} scores (deviation in number of slices) and standard deviation calculated on six anatomical regions of interest of various sizes (cortex, striatum, hippocampus, thalamus, globus pallidus, and substantia nigra) using realistic conventional histological protocols for the six autofluorescence brains (M_1-M_6) and the seven cresyl brains (M_7-M_{13}) in three different contexts: (1) using the giRAff method considering each slice as single, (2) using the giRAff$_m$ extension considering multi-slices and RSF γ_{Mi} evaluated for each brain thanks to the ER_z, and (3) using the giRAff$_m$ extension considering multi-slices and an averaged RSF γ_m evaluated for a given protocol and imaging modality (autofluorescence and cresyl violet).

this improves the result in their opinion for their own dataset, they should obviously reuse it by default for the next iterations with other data produced in the same modality. Initialization of the registration by centering the slices on each other was therefore a mandatory step in this gradual pipeline. Even if this centering process was presented as being manually performed in this study, it would be possible to readily add a simple algorithm to perform this task in an automated manner. Maximizing the overlap of the binarized tissue surface could be used, for example, to improve the method in the future.

The giRAff method was inspired by the operating mode an expert uses when manually identifying the position of a single slice: neuroanatomists flip atlas pages and try to match the shapes of certain anatomical regions in an exploratory way, as well as qualitatively estimate the similarity in a visual manner. Our pipeline does the same using linear registration and NMI. Although NMI has shown its robustness in various multimodal brain applications, its efficiency remains discussed within the scientific community (Zheng, 2006; Xiong et al., 2018; Song et al., 2020). This similarity metric is known to have non-significative values in absolute: comparing two objects whose nature does not have anything in common can even result in a significantly high NMI score (Rohlfing, 2011). The use of NMI was solely relative in our pipeline, considering its score on any template slice in comparison to each other. This information was never used in an absolute manner, and the nature of the objects being compared was the same, thus avoiding this limitation. The NMI was not used as a similarity metric to estimate registration but only to objectively evaluate the quality of the slice-to-slice correspondence after registration.

In the dataset we used, we purposely selected uncut 3D coherent histological brain volumes (autofluorescence of a cleared brain

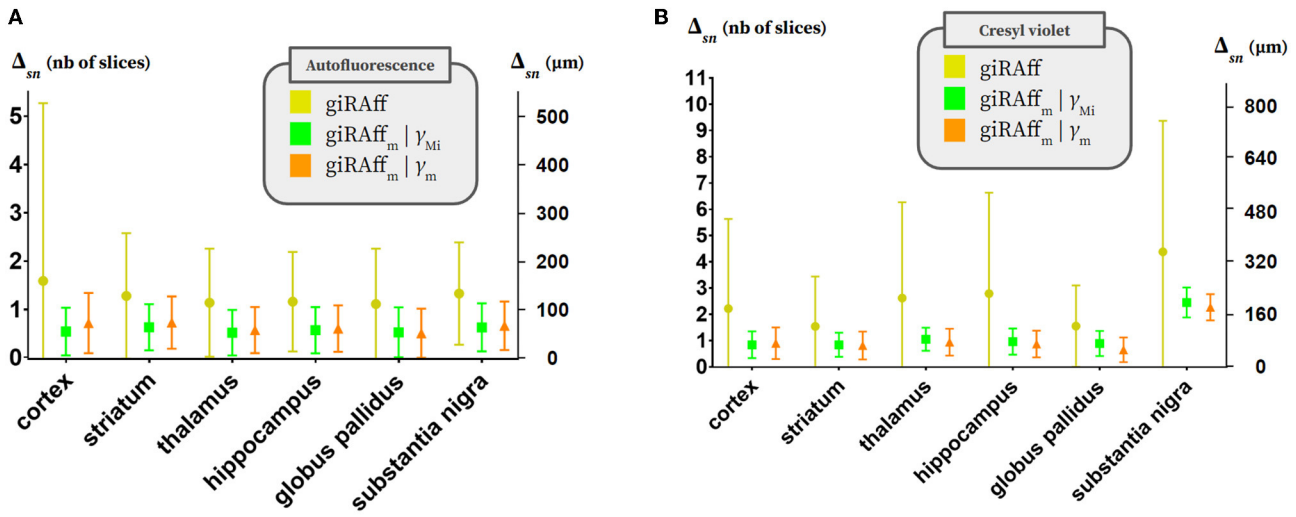


FIGURE 7
Multi-slices. Averaged Δ_{sn} values and standard deviation per considered anatomical region (cortex, striatum, hippocampus, thalamus, globus pallidus, and substantia nigra) in three different contexts: using the giRAff method considering each slice as single (yellow), using the giRAff$_m$ extension considering multi-slices and RSF γ_{Mi} evaluated for each brain thanks to the ER_z (green), and using the giRAff$_m$ extension considering multi-slices and an averaged RSF γ_m evaluated for a given protocol and imaging modality (orange) for **(A)** the six autofluorescence brains and **(B)** the seven cresyl violet brains.

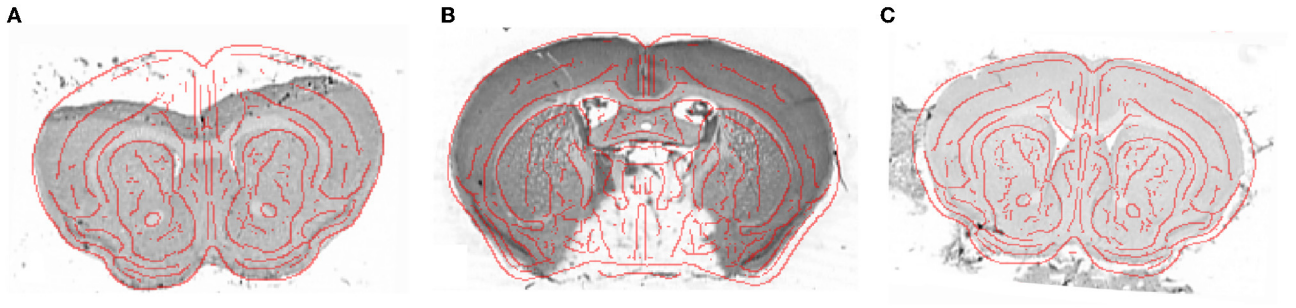


FIGURE 8
Example of histological single slices (cresyl violet) presenting histological artifacts: **(A)** tissue folding (M_{11}), **(B)** missing tissue (M_8), and **(C)** external noise (M_7), with the superposition of the red boundaries of the template slice, obtained using a Deriche filter (Deriche, 1990), at the *z-position* identified by the giRAff$_m$ extension.

acquired with a light-sheet microscope), which was considered as a succession of 2D virtual slices. In such a way, it was possible to test different data processing approaches with artifact-free tissue. This could be one of the reasons why the precision of the *z-position* detection was better for the autofluorescence (lower Δ_{sn}) than for the cresyl violet. We first opted for this favorable context to make a proof of concept, taking autofluorescence as a kind of ideal case (Piluso et al., 2021a). Then, we confronted with "real life" histological preclinical routine protocols (digitized Nissl/cresyl violet-stained brain sections), based on our robust and adjustable pipeline.

When using the giRAff method for a given individual slice, its *z-position* is estimated only once. This estimate may suffer from deviations that could be due to the presence of artifacts in the slices, by poor quality registration, or by a relative similarity value that is not significant enough. In the giRAff$_m$ approach, the joint estimation of the position z_m from a set of slices provided a statistical quantity of estimates sufficient to significantly reduce the deviation Δ_{sn} and its dispersion in general. This improvement was based on the assumption that a large majority of the slices had little or no artifact, and that the registration and similarity metric were robust enough to accurately estimate the *z-position* of such slices in the dataset. As a result, for one or several slice(s) suffering from artifacts, representing a minor proportion of a given dataset, giRAff$_m$ provided better *z-position* identification results than giRAff.

On average, for a multi-slices dataset, the *z-position* of single slices was detected with a precision of one slice in the atlas ($\sim$ 100 µm). This deviation is comparable to the one that experts could make on such a dataset, as long as one single slice considered does not perfectly match one given slice in the template. Indeed, because of its slice thickness and its exact location on the AP axis, as well as possible tilting angles, experts sometimes hesitate between two adjacent slices from the template to identify the right *z-position*

of an experimental single slice. Therefore, they are constrained to make an arbitrary choice, assuming that the position they have identified is only accurate within one slice (100 μm).

Some regions with little pixel support, such as substantia nigra, presented poor dice score results compared to other bigger regions with larger pixel supports. In the linear registration algorithm used, few degrees of liberty were allowed to try to optimize a global transformation at a whole-image scale. This obviously tended to maximize the overlap between regions, including larger pixel support at the expense of other smaller regions including significantly fewer voxels. In such regions, a difference of one single voxel was far more significant than in other regions. Using non-linear registration after estimating the *z-position* could significantly increase the overlap between such small regions and then significantly increase their dice score.

Concerning the cresyl violet data, the M_7 brain showed higher Δ_{sn} scores (larger deviation) than the other brains without using the multi-slices extension giRAff$_m$, confirmed by the presence of artifacts due to the histological and digitization protocols (bubbles, added tissue fragments, and external noise). This is a typical example of artifacts that can occur during a conventional histological protocol. Automatically segmenting histological slices with significant artifacts has always been a challenging task for the scientific community (Agarwal et al., 2016). Most of the time, automatic atlas segmentation of these slices is basically impossible. Our proposed giRAff$_m$ extension has the advantage of optimizing *z-position* detection on a set E of multiple single slices, and thus could be able to identify and segment such slices including artifacts. Results were pooled to obtain the best z_m position estimated for all the slices. Thus, for a set of slices from the same brain, including slices with important artifacts, it was then possible to decrease their rate of contribution π_s (until 0) in the global estimation of the z_m position, but yet achieve their automatic segmentation reliably. Considering a majority of good quality slices selected from E and a robust regression (significantly high coefficient of determination, typically above 0.97), the giRAff$_m$ extension can propose an automated atlas segmentation corresponding for any other slice suffering from those artifacts from the same brain in a robust way, especially without taking them into account in the global estimation of the position z_m. If the rate of contribution π_s was presented as a subjective parameter to add manually as input information within the multi-slices pipeline, further improvements could lead to the use of image processing algorithms able to automatically detect artifacts within histological slices (Agarwal et al., 2016). This would lead to an automated setup of the rate of contribution π_s for each single slice as a consequence. Moreover, using the multi-slices giRAff$_m$ extension allowed for automated estimation of the RSF γ between the data considered. This reinforced the fact the method we proposed is versatile, robust, and adaptable to many types of protocols or histological brain data.

Considering a multi-slices dataset, we focused on a constant inter-slice distance between single slices under study in this article. But in practice, this distance could be heterogeneous. The principle of the multi-slices extension giRAff$_m$ for the analysis of such slices would be exactly the same; the different inter-slice distances can be given as input information within the giRAff$_m$ pipeline.

In conventional histological protocols, tilting angles may occur when slicing the 3D organ. A non-zero φ angle around the IS axis can generate anatomical differences between the left and right side of the slice, which are easy for an expert to identify due to the brain symmetry with respect to the interhemispheric plane. However, it is more challenging to identify a non-zero β angle around the LR axis that will generate differences between the top and bottom of the slice. This angle is most often observed as non-zero, and neurobiologists then have to deal with neighboring slices to perform the segmentation manually. Thanks to a rigid 3D-3D registration between each considered brain and the template volume, it was possible to estimate these tilting angles around the IS and LR axis, and they are of low amplitude ($< 5°$, see Supplementary material S3), hence our focus on the *z-position* determination. Considering those realistic tilting angles of low amplitude, the accuracy of the giRAff method nevertheless made it possible to preserve automatic segmentations for which dice scores are still comparable to those of an expert. Indeed, as protocols for acquiring those brains may be representative of standard protocols in conventional brain histology performed in the coronal incidence, we assume that tilting angles rarely exceed an amplitude of 5° with modern equipment and in a similar study framework. If this angulation generates genuine anatomical differences compared to data without angulation, the method we proposed made it possible to compensate for this drawback. Indeed, we chose to process data produced in routine histological protocols in this article, i.e., including real tilting angles caused by the cutting process. Histological data presented in this article included their native tilting angles. As the giRAff method detected the *z-position* of the single slice with high accuracy, its anatomical environment was well identified (basically in the thickness range of about 200 μm). Following this location, registration ensured the best matching of the tissue between the single slice considered and the template slice identified, as it would have been done in the case of considering the respective adjacent slices of its direct neighborhood. In the coronal incidence and with a slice thickness of about a hundred micrometers, anatomical variations are small from one slice to the next adjacent one. The template data are smooth, and very few discontinuities appear when examining the slices one after the other along the AP axis. More specifically, a tilting angle would generate small anatomical differences between the right and left of the slice for an angle φ around the IS axis and between the top and bottom of the slice for an angle β around the LR axis compared to the template data. In practice, using linear registration would basically correct most of those segmentation errors because the presence and location of anatomical regions are almost the same from one slice n to its n-1 and n+1 (or more) neighbors. Indeed, anatomical differences generated by a tilting angle cause linear deformations along one, two, or both axes (IS and LR in the coronal incidence), which affine registration can compensate with shearing. This was confirmed by dice scores calculated, which were widely comparable to those of an expert in the end. The only necessary condition is that the *z-position* of the single slice considered must be accurately estimated, typically with a deviation less or equal to one slice in the template, to avoid too large anatomical difference between slices considered. It is just a matter of comparing data which are

comparable, i.e., extracted, cut, and digitized within a rigorous, consistent, and realistic study framework. If neurobiologists are asked to cut coronal mouse brain slices using a microtome, it is reasonable to believe that their skills will enable them to obtain tilting angles below 5° as observed in the data presented in this article. Visual quality control and steel matrices could also be used for this purpose.

The method we proposed was based on linear registration in a pipeline with an increasing number of degrees of freedom. The use of non-linear registration could compromise the identification of the correct *z-position* of a given single slice. Indeed, too many degrees of freedom would excessively distort all template slices to match, in an inappropriate way, the single slice under consideration. It would then be challenging to distinguish which was the most similar. In contrast, the use of non-linear 2D-2D registration between the single slice and the template slice identified at the *z-position* at the end of the giRAff pipeline would certainly enable the segmentation results to be refined. This could be useful for the analysis of small regions, for example. This constitutes one of the further improvements the method could benefit from. Moreover, the lack of ground truth will make the task even harder.

A benchmark between the different methods of segmenting single slices should be carried out to identify which could give the best results according to the experimental data under study. Such a benchmark should accurately compare all the methods using a dedicated common dataset as well as an appropriate metric to evaluate their respective performance. This comparison is too vast to be presented exhaustively and precisely in this paper and could be the scope of another study. Indeed, each method has its own particular way of working, and its results may be of a different nature, making them difficult to benchmark. Nevertheless, we wanted to briefly test whether our method offered competitive results compared with those provided by the most recent state-of-the-art method. A quick comparison was led on two independent single coronal Nissl-stained slices between the latest fully automated method from the state-of-the-art (DeepSlice from Carey et al., 2023) and our giRAff method. We estimated NMI similarity metric after applying both methods in the same conditions. Those unitary tests showed that similarity between the resulting slices from our method outperformed DeepSlice by about 20%, while requiring a longer processing time ($<$ 30 s for DeepSlice and about 1 min for giRAff, estimated per slice). Looking at the anatomy in the identified template slices, the *z-position* determined for both methods was very close, if not equal. Only some slight registration differences were observed, where the registration algorithm used in giRAff provided the best results according to the NMI criterion. These were very preliminary unitary tests, hence the need for this benchmark to be fully explored in future.

The giRAff method was developed to be fully automatic and embedded in an easy-to-use interface with very few input parameters so that it can be easily used by a non-expert. Optional parameters can be adjusted if the user wants to contribute with their own knowledge, such as the selection of the region(s) of interest studied. This information will reduce the number of adjacent template slices to consider in estimating the *z-position* of a single slice. Only template slices including this or those anatomical region(s) will be pre-selected, thus decreasing the computation time.

In automatic mode, the method segments a single slice in 1 min on a high-performance computing infrastructure. The result benefits not only from the six regions we focused on but from all the subregions defined in the ABA reference. This is comparable to the time it may take an expert to identify the correct *z-position* of a single slice within the atlas template. For the same processing time, the giRAff method additionally provided direct atlas segmentation of the single slice. Moreover, no knowledge of brain anatomy or even in coding was required to use the method. Its interface and the few input parameters required by our pipeline make it usable by anyone with full autonomy. Even without supercomputer infrastructure, using about 20 computing cores from a workstation, for example, the method for one single slice worked in a reasonable time of about 15 min.

First, preliminary results as well as complementary studies on a brain suffering from pathological lesions showed encouraging results for the method to be able to handle such data in the context of dedicated protocols. This opens the door for automated segmentation of slices from pathological mouse models, whether neurodegenerative or other diseases, as long as data did not suffer from too large anatomical alterations. Similarly, the use of this pipeline can be extended to other rodents, such as rat for instance, or even in other modalities, such as magnetic resonance imaging. Promising results have been obtained on this modality (Piluso et al., 2021b), and future work aims at validating the use of the method in such cases. In addition, the use of this method will indirectly allow better targeting of conventional histology protocols to reduce the amount of brain data to be used in a study.

5 Conclusion

The wide variety of existing histological protocols as well as the great numbers of anatomical structures in the mouse brain makes the analysis of histological slices quite tedious and complex. In conventional preclinical histology for the analysis of the mouse brain, it is rare to have enough slices to reconstruct the brain in 3D and, sometimes, working on 3D data is not a prerequisite. It is possible to study only one single slice within the brain, but this is also unusual. In contrast, many protocols are based on a fairly large number of slices to perform quantitative studies on particular anatomical regions or around a specific pathological lesion, for example, still precluding 3D reconstruction. Whatever the case, the generic giRAffMapper pipeline was optimized to accommodate most protocols involving any number of single slices. We showed that our method was able to automatically identify the position of single slices within a mouse brain atlas with less than one slice deviation on average and in 1 min for one slice. Atlas segmentations were comparable to those of an expert. The giRAff method does not need any 3D brain volume reconstruction; it is versatile, generic, user-friendly, and requires few input parameters. In future, we aim to take into account real slice angles and use non-linear registration tools to further refine the segmentation of anatomical regions from increasingly precise atlases. This study paves the way for automated atlas segmentation through a simplified interface of any histological mouse slice, half- or whole-brain slice, for pathological models,

for different modalities and possibly for different species. This is done in a fully automated way and does not require any particular knowledge of the study involved, nor in neuroanatomy in general, nor even in coding, to be able to use it. This significantly widens the scope of use of such anatomical detailed atlases within the scientific community for a complex task that usually had to be performed only by experts.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The animal study was approved by dedicated Institutional Animal Care and Ethics Committees, where the experimental procedures involving animal models described in this paper come from already published papers (Renier et al., 2016; Vandenberghe et al., 2016). The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

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

Funding

This study was supported by Association Nationale Recherche Technologie 2018/0809.

Conflict of interest

SP and CC were employed by WITSEE.

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.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2023.1230814/full#supplementary-material

References

Agarwal, N., Xu, X., and Gopi, M. (2016). "Automatic detection of histological artifacts in mouse brain slice images," in *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging* (Cham: Springer), 105–115. doi: 10.1007/978-3-319-61188-4_10

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2018). "An unsupervised learning model for deformable medical image registration," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 9252–9260. doi: 10.1109/CVPR.2018.00964

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* 38, 1788–1800. doi: 10.1109/TMI.2019.2897538

Bayraktar, O. A., Bartels, T., Holmqvist, S., Kleshchevnikov, V., Martirosyan, A., Polioudakis, D., et al. (2020). Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. *Nat. Neurosci.* 23, 500–509. doi: 10.1038/s41593-020-0602-1

Berlanga, M. L., Phan, S., Bushong, E. A., Wu, S., Kwon, O., Phung, B. S., et al. (2011). Three-dimensional reconstruction of serial mouse brain sections: solution for flattening high-resolution large-scale mosaics. *Front. Neuroanat.* 5, 17. doi: 10.3389/fnana.2011.00017

Bohland, J. W., Bokil, H., Pathak, S. D., Lee, C. K., Ng, L., Lau, C., et al. (2010). Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods* 50, 105–112. doi: 10.1016/j.ymeth.2009.09.001

Borovec, J., Kybic, J., Arganda-Carreras, I., Sorokin, D. V., Bueno, G., Khvostikov, A. V., et al. (2020). ANHIR: automatic non-rigid histological image registration challenge. *IEEE Trans. Med. Imaging* 39, 3042–3052. doi: 10.1109/TMI.2020.2986331

Carey, H., Pegios, M., Martin, L., Saleeba, C., Turner, A. J., Everett, N. A., et al. (2023). DeepSlice: rapid fully automatic registration of mouse brain imaging to a volumetric atlas. *Nat. Commun.* 14, 5884. doi: 10.1038/s41467-023-41645-4

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

Chon, U., Vanselow, D. J., Cheng, K. C., and Kim, Y. (2019). Enhanced and unified anatomical labeling for a common mouse brain atlas. *Nat. Commun.* 10, 1–12. doi: 10.1038/s41467-019-13057-w

Cointepas, Y., Mangin, J. F., Garnero, L., Poline, J. B., and Benali, H. (2001). BrainVISA: software platform for visualization and analysis of multi-modality brain data. *Neuroimage* 13, 98. doi: 10.1016/S1053-8119(01)91441-7

Costa, M., Manton, J. D., Ostrovsky, A. D., Prohaska, S., and Jefferis, G. S. (2016). NBLAST: rapid, sensitive comparison of neuronal structure and construction of neuron family databases. *Neuron* 91, 293–311. doi: 10.1016/j.neuron.2016.06.012

Dauguet, J., Delzescaux, T., Condé, F., Mangin, J. F., Ayache, N., Hantraye, P., et al. (2007). Three-dimensional reconstruction of stained histological slices and 3D non-linear registration with in-vivo MRI for whole baboon brain. *J. Neurosci. Methods* 164, 191–204. doi: 10.1016/j.jneumeth.2007.04.017

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., and Išgum, I. (2019). A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* 52, 128–143. doi: 10.1016/j.media.2018.11.010

de Vos, B. D. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I. (2017). "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Cham: Springer), 204–212. doi: 10.1007/978-3-319-67558-9_24

Deriche, R. (1990). Fast algorithms for low-level vision. *IEEE Trans. Pattern Anal. Mach. Intell.* 12, 78–87. doi: 10.1109/34.41386

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

Dorocic, I. P., Fürth, D., Xuan, Y., Johansson, Y., Pozzi, L., Silberberg, G., et al. (2014). A whole-brain atlas of inputs to serotonergic neurons of the dorsal and median raphe nuclei. *Neuron* 83, 663–678. doi: 10.1016/j.neuron.2014.07.002

Dostrovsky, J. O., Hutchison, W. D., and Lozano, A. M. (2002). The globus pallidus, deep brain stimulation, and Parkinson's disease. *Neuroscientist* 8, 284–290. doi: 10.1177/1073858402008003014

Dubois, A., Hérard, A. S., Delatour, B., Hantraye, P., Bonvento, G., Dhenain, M., et al. (2010). Detection by voxel-wise statistical analysis of significant changes in regional cerebral glucose uptake in an APP/PS1 transgenic mouse model of Alzheimer's disease. *Neuroimage* 51, 586–598. doi: 10.1016/j.neuroimage.2010.02.074

Dudeffant, C., Vandesquille, M., Herbert, K., Garin, C. M., Alves, S., Blanchard, V., et al. (2017). Contrast-enhanced MR microscopy of amyloid plaques in five mouse models of amyloidosis and in human Alzheimer's disease brains. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-05285-1

Eastwood, B. S., Hooks, B. M., Paletzki, R. F., O'Connor, N. J., Glaser, J. R., and Gerfen, C. R. (2019). Whole mouse brain reconstruction and registration to a reference atlas with standard histochemical processing of coronal sections. *J. Comp. Neurol.* 527, 2170–2178. doi: 10.1002/cne.24602

Erö, C., Gewaltig, M. O., Keller, D., and Markram, H. (2018). A cell atlas for the mouse brain. *Front. Neuroinform.* 12:84. doi: 10.3389/fninf.2018.00084

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., et al. (2019). U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* 16, 67–70. doi: 10.1038/s41592-018-0261-2

Fürth, D., Vaissière, T., Tzortzi, O., Xuan, Y., Märtin, A., Lazaridis, I., et al. (2018). An interactive framework for whole-brain maps at cellular resolution. *Nat. Neurosci.* 21, 139–149. doi: 10.1038/s41593-017-0027-7

Geha, P. Y., Baliki, M. N., Harden, R. N., Bauer, W. R., Parrish, T. B., and Apkarian, A. V. (2008). The brain in chronic CRPS pain: abnormal gray-white matter interactions in emotional and autonomic regions. *Neuron* 60, 570–581. doi: 10.1016/j.neuron.2008.08.022

Ghaznavi, F., Evans, A., Madabhushi, A., and Feldman, M. (2013). Digital imaging in pathology: whole-slide imaging and beyond. *Annu. Rev. Pathol.* 8, 331–359. doi: 10.1146/annurev-pathol-011811-120902

Henderson, M. X., Cornblath, E. J., Darwich, A., Zhang, B., Brown, H., Gathagan, R. J., et al. (2019). Spread of α-synuclein pathology through the brain connectome is modulated by selective vulnerability and predicted by network analysis. *Nat. Neurosci.* 22, 1248–1257. doi: 10.1038/s41593-019-0457-5

Hérard, A. S., Petit, F., Gary, C., Guillermier, M., Boluda, S., Garin, C. M., et al. (2020). Induction of amyloid-β deposits from serially transmitted, histologically silent, Aβ seeds issued from human brains. *Acta Neuropathol. Commun.* 8, 1–10. doi: 10.1186/s40478-020-01081-7

Iglesias, J. E., Modat, M., Peter, L., Stevens, A., Annunziata, R., Vercauteren, T., et al. (2018). Joint registration and synthesis using a probabilistic model for alignment of MRI and histological sections. *Med. Image Anal.* 50, 127–144. doi: 10.1016/j.media.2018.09.002

Jefferis, G. S., Potter, C. J., Chan, A. M., Marin, E. C., Rohlfing, T., Maurer Jr, C. R., et al. (2007). Comprehensive maps of Drosophila higher olfactory centers: spatially segregated fruit and pheromone representation. *Cell* 128, 1187–1203. doi: 10.1016/j.cell.2007.01.040

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

Kim, S. Y., Cho, J. H., Murray, E., Bakh, N., Choi, H., Ohn, K., et al. (2015). Stochastic electrotransport selectively enhances the transport of highly electromobile molecules. *Proc. Nat. Acad. Sci.* 112, E6274–E6283. doi: 10.1073/pnas.1510133112

Kim, Y., Yang, G. R., Pradhan, K., Venkataraju, K. U., Bota, M., Del Molino, L. C. G., et al. (2017). Brain-wide maps reveal stereotyped cell-type-based cortical architecture and subcortical sexual dimorphism. *Cell* 171, 456–469. doi: 10.1016/j.cell.2017.09.020

Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F. C., Miao, S., et al. (2017). "Robust non-rigid registration through agent-based action learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Cham: Springer), 344–352. doi: 10.1007/978-3-319-66182-7_40

Krepl, J., Casalegno, F., Delattre, E., Erö, C., Lu, H., Keller, D., et al. (2021). Supervised learning with perceptual similarity for multimodal gene expression registration of a mouse brain atlas. *Front. Neuroinform.* 15:691918. doi: 10.3389/fninf.2021.691918

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

Laguitton, S., Riviere, D., Vincent, T., Fischer, C., Geffroy, D., Souedet, N., et al. (2011). "Soma-workflow: a unified and simple interface to parallel computing resources," in *MICCAI Workshop on High Performance and Distributed Computing for Medical Imaging*, eds G. Fichtinger, A. Martel, and T. Peters (Berlin; Heidelberg: Springer-Verlag).

Lam, S., Hérard, A. S., Boluda, S., Petit, F., Eddarkaoui, S., Cambon, K., et al. (2022). Pathological changes induced by Alzheimer's brain inoculation in amyloid-beta plaque-bearing mice. *Acta Neuropathol. Commun.* 10, 1–19. doi: 10.1186/s40478-022-01410-y

Lau, C., Ng, L., Thompson, C., Pathak, S., Kuan, L., Jones, A., et al. (2008). Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics* 9:153. doi: 10.1186/1471-2105-9-153

Lebenberg, J., Hérard, A. S., Dubois, A., Dauguet, J., Frouin, V., Dhenain, M., et al. (2010). Validation of MRI-based 3D digital atlas registration with histological and autoradiographic volumes: an anatomofunctional transgenic mouse brain imaging study. *Neuroimage* 51, 1037–1046. doi: 10.1016/j.neuroimage.2010.03.014

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176. doi: 10.1038/nature05453

Li, H., and Fan, Y. (2017). Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv preprint arXiv*. doi: 10.1109/ISBI.2018.8363757

Mancini, M., Casamitjana, A., Peter, L., Robinson, E., Crampsie, S., Thomas, D. L., et al. (2020). A multimodal computational pipeline for 3D histology of the human brain. *Sci. Rep.* 10, 1–21. doi: 10.1038/s41598-020-69163-z

Mesejo, P., Ugolotti, R., Cagnoni, S., Di Cunto, F., and Giacobini, M. (2012). "Automatic segmentation of hippocampus in histological images of mouse brains using deformable models and random forest," in *2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS)* (Rome: IEEE), 1–4. doi: 10.1109/CBMS.2012.6266318

Milligan, K., Balwani, A., and Dyer, E. (2019). Brain mapping at high resolutions: challenges and opportunities. *Curr. Opin. Biomed. Eng.* 12, 126–131. doi: 10.1016/j.cobme.2019.10.009

Modat, M., Cash, D. M., Daga, P., Winston, G. P., Duncan, J. S., and Ourselin, S. (2014). Global image registration using a symmetric block-matching approach. *J. Med. Imag.* 1, 024003. doi: 10.1117/1.JMI.1.2.024003

Niedworok, C. J., Brown, A. P., Cardoso, M. J., Osten, P., Ourselin, S., Modat, M., et al. (2016). aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. *Nat. Commun.* 7, 1–9. doi: 10.1038/ncomms11879

Ourselin, S., Roche, A., Subsol, G., Pennec, X., and Ayache, N. (2001). Reconstructing a 3D structure from serial histological sections. *Image Vis. Comput.* 19, 25–31. doi: 10.1016/S0262-8856(00)00052-4

Pagani, M., Damiano, M., Galbusera, A., Tsaftaris, S. A., and Gozzi, A. (2016). Semi-automated registration-based anatomical labelling, voxel based morphometry and cortical thickness mapping of the mouse brain. *J. Neurosci. Methods* 267, 62–73. doi: 10.1016/j.jneumeth.2016.04.007

Pallast, N., Wieters, F., Fink, G. R., and Aswendt, M. (2019). Atlas-based imaging data analysis tool for quantitative mouse brain histology (AIDAhisto). *J. Neurosci. Methods* 326:108394. doi: 10.1016/j.jneumeth.2019.108394

Papp, E. A., Leergaard, T. B., Calabrese, E., Johnson, G. A., and Bjaalie, J. G. (2014). Waxholm space atlas of the sprague dawley rat brain. *Neuroimage* 97, 374–386. doi: 10.1016/j.neuroimage.2014.04.001

Picconi, B., Pisani, A., Barone, I., Bonsi, P., Centonze, D., Bernardi, G., et al. (2005). Pathological synaptic plasticity in the striatum: implications for Parkinson's disease. *Neurotoxicology* 26, 779–783. doi: 10.1016/j.neuro.2005.02.002

Pichat, J., Iglesias, J. E., Yousry, T., Ourselin, S., and Modat, M. (2018). A survey of methods for 3D histology reconstruction. *Med. Image Anal.* 46, 73–105. doi: 10.1016/j.media.2018.02.004

Piluso, S., Souedet, N., Flament, J., Gaudin, M., Jan, C., Bonvento, G., et al. (2021b). "Automated correspondence and registration between CEST and histological mouse brain sections for in vivo virus tracking" in *2021 European Society for Magnetic Resonance in Medicine and Biology (ESMRMB). S6.O7., p. S47. Book of Abstracts ESMRMB 2021 Online 38th Annual Scientific Meeting 7–9 October 2021* (Berlin: Springer).

Piluso, S., Souedet, N., Jan, C., Clouchoux, C., and Delzescaux, T. (2021a). "Automated atlas-based segmentation of single coronal mouse brain slices using linear 2D-2D registration," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Mexico: IEEE), 2860–2863. doi: 10.1109/EMBC46164.2021.9631097

Puchades, M. A., Csucs, G., Ledergerber, D., Leergaard, T. B., and Bjaalie, J. G. (2019). Spatial registration of serial microscopic brain images to three-dimensional reference atlases with the QuickNII tool. *PLoS ONE* 14, e0216796. doi: 10.1371/journal.pone.0216796

Renier, N., Adams, E. L., Kirst, C., Wu, Z., Azevedo, R., Kohl, J., et al. (2016). Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* 165, 1789–1802. doi: 10.1016/j.cell.2016.05.007

Rodarie, D., Verasztó, C., Roussel, Y., Reimann, M., Keller, D., Ramaswamy, S., et al. (2022). A method to estimate the cellular composition of the mouse brain from heterogeneous datasets. *PLOS Comput. Biol.* 18, e1010739. doi: 10.1371/journal.pcbi.1010739

Rohé, M. M., Datar, M., Heimann, T., Sermesant, M., and Pennec, X. (2017). "SVF-Net: Learning deformable image registration using shape matching," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Cham: Springer), 266–274. doi: 10.1007/978-3-319-66182-7_31

Rohlfing, T. (2011). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* 31, 153–163. doi: 10.1109/TMI.2011.2163944

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 (MICCAI)* (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Sadeghi, M., Neto, P., Ramos-Prats, A., Castaldi, F., Paradiso, E., Mahmoodian, N., et al. (2022). "Automatic 2D to 3D localization of histological mouse brain sections in the reference atlas using deep learning," in *Medical Imaging 2022: Image Processing* (San Diego, CA: SPIE), 718–724. doi: 10.1117/12.2604231

Sen, M. K., Almuslehi, M. S., Coorssen, J. R., Mahns, D. A., and Shortland, P. J. (2020). Behavioural and histological changes in cuprizone-fed mice. *Brain Behav. Immun.* 87, 508–523. doi: 10.1016/j.bbi.2020.01.021

Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., and Staring, M. (2017). "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Cham: Springer), 232–239. doi: 10.1007/978-3-319-66182-7_27

Song, J. H., Choi, W., Song, Y. H., Kim, J. H., Jeong, D., Lee, S. H., et al. (2020). Precise Mapping of single neurons by calibrated 3D reconstruction of brain slices reveals topographic projection in mouse visual cortex. *Cell Rep.* 31, 107682. doi: 10.1016/j.celrep.2020.107682

Stæger, F. F., Mortensen, K. N., Nielsen, M. S. N., Sigurdsson, B., Kaufmann, L. K., Hirase, H., et al. (2020). A three-dimensional, population-based average of the C57BL/6 mouse brain from DAPI-stained coronal slices. *Sci. Data* 7, 1–7. doi: 10.1038/s41597-020-0570-z

Stolp, H. B., Ball, G., So, P. W., Tournier, J. D., Jones, M., Thornton, C., et al. (2018). Voxel-wise comparisons of cellular microstructure and diffusion-MRI in mouse hippocampus using 3D Bridging of Optically-clear histology with Neuroimaging Data (3D-BOND). *Sci. Rep.* 8, 1–12. doi: 10.1038/s41598-018-22295-9

Studholme, C., Hawkes, D. J., and Hill, D. L. (1998). "Normalized entropy measure for multimodality image alignment," in *Medical Imaging 1998: Image Processing* (San Diego, CA: International Society for Optics and Photonics), 132–143. doi: 10.1117/12.310835

Tappan, S. J., Eastwood, B. S., O'Connor, N., Wang, Q., Ng, L., Feng, D., et al. (2019). Automatic navigation system for the mouse brain. *J. Comp. Neurol.* 527, 2200–2211. doi: 10.1002/cne.24635

Teichmann, M., Dupoux, E., Kouider, S., Brugières, P., Boissé, M. F., Baudic, S., et al. (2005). The role of the striatum in rule application: the model of Huntington's disease at early stage. *Brain* 128, 1155–1167. doi: 10.1093/brain/awh472

Toga, A. W., Ambach, K., Quinn, B., Hutchin, M., and Burton, J. S. (1994). Postmortem anatomy from cryosectioned whole human brain. *J. Neurosci. Methods* 54, 239–252. doi: 10.1016/0165-0270(94)90196-1

Tward, D. J., Li, X., Huo, B., Lee, B. C., Miller, M., and Mitra, P. P. (2020). Solving the where problem in neuroanatomy: a generative framework with learned mappings to register multimodal, incomplete data into a reference brain. *bioRxiv*. doi: 10.1101/2020.03.22.002618

Vandenberghe, M. E., Hérard, A. S., Souedet, N., Sadouni, E., Santin, M. D., Briet, D., et al. (2016). High-throughput 3D whole-brain quantitative histopathology in rodents. *Sci. Rep.* 6, 20958. doi: 10.1038/srep20958

Wang, Q., Ding, S. L., Li, Y., Royall, J., Feng, D., Lesnar, P., et al. (2020). The allen mouse brain common coordinate framework: a 3D reference atlas. *Cell* 1814, 936–953. doi: 10.1016/j.cell.2020.04.007

Wilt, B. A., Burns, L. D., Wei Ho, E. T., Ghosh, K. K., Mukamel, E. A., and Schnitzer, M. J. (2009). Advances in light microscopy for neuroscience. *Annu. Rev. Neurosci.* 32, 435–506. doi: 10.1146/annurev.neuro.051508.135540

Wu, T., Dejanovic, B., Gandham, V. D., Gogineni, A., Edmonds, R., Schauer, S., et al. (2019). Complement C3 is activated in human AD brain and is required for neurodegeneration in mouse models of amyloidosis and tauopathy. *Cell Rep.* 28, 2111–2123. doi: 10.1016/j.celrep.2019.07.060

Xiong, J., Ren, J., Luo, L., and Horowitz, M. (2018). Mapping histological slice sequences to the allen mouse brain atlas without 3D reconstruction. *Front. Neuroinform.* 12:93. doi: 10.3389/fninf.2018.00093

Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: fast predictive image registration–a deep learning approach. *Neuroimage* 158, 378–396. doi: 10.1016/j.neuroimage.2017.07.008

Yates, S. C., Groeneboom, N. E., Coello, C., Lichtenthaler, S. F., Kuhn, P. H., Demuth, H. U., et al. (2019). QUINT: workflow for quantification and spatial analysis of features in histological images from rodent brain. *Front. Neuroinform.* 13, 75. doi: 10.3389/fninf.2019.00075

Ye, L., Allen, W. E., Thompson, K. R., Tian, Q., Hsueh, B., Ramakrishnan, C., et al. (2016). Wiring and molecular features of prefrontal ensembles representing distinct experiences. *Cell* 165, 1776–1788. doi: 10.1016/j.cell.2016.05.010

Yee, Y., Ellegood, J., French, L., and Lerch, J. P. (2022). Organization of thalamocortical structural covariance and a corresponding 3D atlas of the mouse thalamus. *bioRxiv*. doi: 10.1101/2022.03.10.483857

Zeng, H. (2018). Mesoscale connectomics. *Curr. Opin. Neurobiol.* 50, 154–162. doi: 10.1016/j.conb.2018.03.003

Zheng, G. (2006). "A novel 3D/2D correspondence building method for anatomy-based registration," in *International Workshop on Biomedical Image Registration (WBIR)* (Berlin, Heidelberg: Springer), 75–83. doi: 10.1007/11784012_10

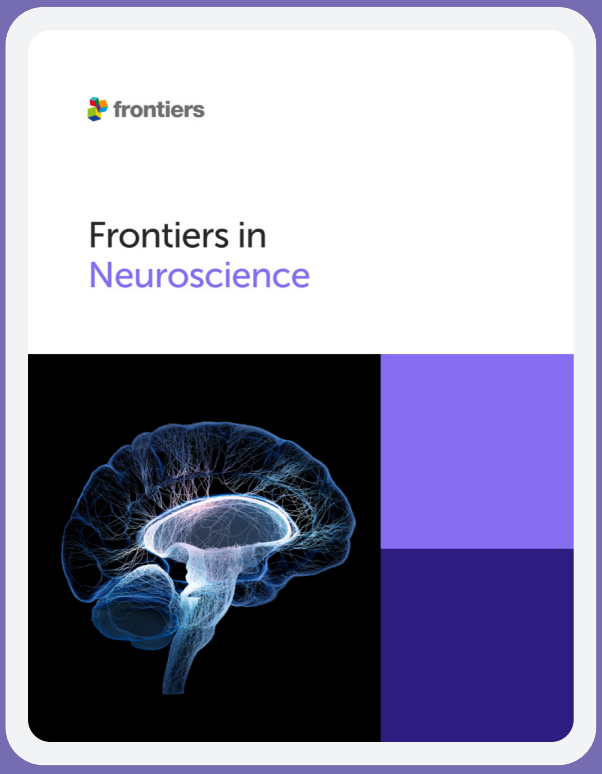
Frontiers in Neuroscience

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

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact