4DfCF: 4D fMRI CrossFormer Vision Transformer

Chensheng Zheng, Shaker El-Sappagh, and Tamer Abuhmed, *Senior Member, IEEE*

***Abstract*—Investigating the spatiotemporal dynamics of the human brain is a complex challenge due to the intricate nature of brain networks, and the limitations of current analytical methods. Herein, we introduce the 4D functional Magnetic Resonance Imaging (fMRI) CrossFormer (4DfCF), a novel vision transformer architecture designed to process high-dimensional 4D fMRI data. This model integrates temporal and spatial dimensions to effectively learn and predict cognitive and clinical outcomes. We further evaluated the 4DfCF on three benchmark datasets: Attention Deficit Hyperactivity Disorder-200 (ADHD-200), Alzheimer's Disease Neuroimaging Initiative (ADNI), and Autism Brain Imaging Data Exchange (ABIDE). The results showed that our model consistently outperforms state-of-the-art baseline models, achieving an accuracy improvement of 5–10% , a precision increase of 4–8% , a recall enhancement of 6–9% , and an F1-score boost of 7–11% . Additionally, the 4D fMRI CrossFormer-Tiny variant demonstrated greater efficiency than existing methods, using 20% fewer computational resources and achieving 30% faster training times. Pre-training experiments further reveal that models pre-trained on one dataset and fine-tuned on another achieved faster convergence and higher accuracy, with the Autism Brain Imaging Data Exchange (ABIDE) pre-trained models showing the best performance. Additionally, we employed an explainable AI method to identify the brain regions associated with disease diagnosis. Overall, our findings highlight the potential of the 4DfCF to advance precision neuroscience through efficient and scalable analysis of complex fMRI data.**

***Index Terms*—4D fMRI, brain disorders, spatiotemporal dynamics, vision transformer.**

I. Introduction

FUNCTIONAL Magnetic Resonance Imaging (fMRI) is a powerful tool used to measure and map brain activity. It captures the temporal sequence of 3D images of Blood Oxygenation Level-Dependent (BOLD) signals. As such, fMRI provides a unique opportunity to model the spatiotemporal patterns of brain electrochemical activity. This ability to visualize brain function in real-time makes fMRI a critical resource for understanding both normal and pathological brain processes [1]. However, the complexity of the human brain induces significant challenges in fMRI analysis. The intricate nature of brain networks and the high dimensionality of fMRI data complicate the effective modeling and interpretation of brain activity [2], [3]. Traditional statistical methods and many advanced deep learning approaches have struggled to capture the nonlinear and dynamic relationships inherent in brain functions [4], [5], [6]. These challenges necessitate the development of more sophisticated models capable of handling the complexity of fMRI data [7]. Two distinct Deep Neural Network (DNN)-based approaches, can achieve this. The first is a Region of Interest (ROI)-based method in which high-dimensional fMRI data are clustered into temporal sequences of hundreds of predefined brain regions (i.e., ROIs) before learning a predictive DNN model [8]. Although computationally efficient, this pre-processing can lose crucial information required to capture subtle variability across individual brains [9]. The second approach is the two-step method, in which raw fMRI data are used as input, and specialized architectures for the spatial and temporal domains are separately employed [10]. Convolutional Neural Networks (CNNs) are used for spatial features, while Long Short-Term Memory (LSTM) or transformers are used for temporal features [11], [12]. This separation enhances the model's computational and memory efficiency, but may limit the capability of capturing comprehensive information among brain regions across the temporal dimension [13]. A review of the existing literature revealed that the development of 4D fMRI models has seen significant advancements with the introduction of models such as the Swim 4D fMRI transformer (SwiFT) [14], which employs a 4D local window attention structure to focus on local regions of fMRI volumes, capturing fine-grained spatiotemporal patterns within small neighborhoods. However, SwiFT has some limitations, particularly in capturing long-range dependencies across the entire fMRI volume, due to its localized attention mechanism. Therefore, to address these limitations, we propose the 4D fMRI CrossFormer (4DfCF), a novel vision transformer designed to process 4D fMRI data. We hypothesized that integrating

Received 11 March 2025; revised 8 August 2025 and 9 October 2025; accepted 7 November 2025. Date of publication 13 November 2025; date of current version 5 June 2026. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant 2021R1A2C1011198, in part by the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) through the ICT Creative Consilience Program under Grant IITP-2021-2020-0-01821, and in part by AI Platform to Fully Adapt and Reflect Privacy-Policy Changes under Grant 2022-0-00688. *(Corresponding author: Tamer Abuhmed.)*

Chensheng Zheng and Tamer Abuhmed are with the College of Computing and Informatics, Sungkyunkwan University, Suwon 16419, South Korea (e-mail: chriszheng07@g.skku.edu; tamer@skku.edu).

Shaker El-Sappagh is with the Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt (e-mail: shaker.elsappagh@gu.edu.eg).

The code is publicly available at: https://github.com/InfoLab-SKKU/4DfCF.

Digital Object Identifier 10.1109/JBHI.2025.3631655

temporal and spatial dimensions within a single model would allow for more accurate and efficient learning of brain dynamics, thereby achieving the better prediction of cognitive and clinical outcomes. The 4DfCF integrates both local and global attention mechanisms to capture intricate spatiotemporal relationships. Compared with SwiFT, which employs only local window attention, 4DfCF unifies multiresolution patch embeddings with hierarchical 4D short and longdistance attention, enabling simultaneous capture of finegrained and global spatiotemporal dependencies. This design overcomes SwiFT's inability to model longrange volumewide interactions, preserves full volumetric structure, and scales linearly with data size, yielding more robust and accurate clinical predictions. The model effectively processed and learned from high-dimensional 4D fMRI data, achieving significant performance improvements over current state-of-the-art models. By addressing these limitations, we aimed to advance precision neuroscience through more accurate and efficient brain dynamics analysis.

The main contributions of our work are as follows: (1) we propose a novel transformer architecture, 4DfCF, integrating temporal and spatial dimensions to handle the complexity of 4D fMRI data, and (2) we demonstrate that our model significantly outperforms existing methods in terms of accuracy, precision, recall, and F1-score across multiple benchmark datasets: ADHD-200 [22], ADNI [23], and ABIDE [15]; (3) we show that the 4D fMRI CrossFormer-Tiny (4DfCF-T) variant achieves greater computational efficiency, using fewer resources and achieving faster training times; (4) we validate the effectiveness of pre-training on one dataset and fine-tuning on another, highlighting the potential to build large-scale foundation models for fMRI analysis; (5) we present the explainable results using an Integrated Gradient with Smoothgrad sQuare (IG-SQ) [16] for 4DfCF's predictions and conduct ablation studies to substantiate our modeling choices. By addressing the complexity of fMRI data using 4DfCF, we aimed to advance the field of precision neuroscience by providing a robust and scalable solution for the analysis of complex brain imaging data. The remainder of this paper is organized as follows: Section II details the overall architecture of the proposed 4DfCF, including the 4D Cross-Scale Embedding Layer and the 4D CrossFormer Blocks. Section III describes the experimental settings, implementation details, performance evaluation, and discussion on benchmark datasets; Section IV indicates the clinical and technical impact of the proposed work and comparison with SOTA; and Section V summarizes the findings and discusses potential future research directions.

II. METHODS

The proposed 4DfCF is based on CrossFormer [17], which was modified from its original design to handle 2D images. Our modifications enable this software to process 4D fMRI images, making it suitable for complex brain disease classification tasks. Specifically, the pre-processing steps included standard corrections for slice-timing differences, motion artifacts, physiological noise, and temporal drifts. Additionally, the datasets were aligned with the anatomical scans and transformed into a standardized space. The volumes were resampled and smoothed to enhance data quality. These pre-processing steps ensured the suitability of the data for analysis. fMRI images were used to train the 4DfCF model. Finally, based on the learned latent representations, a specialized head (*e.g.*, the classification head in Fig. 1(a)) was integrated to classify the fMRI images and diagnose brain diseases. Fig. 1 illustrates the detailed architecture of the proposed 4DfCF. In the following sections, we present the components of the 4DfCF (Section II-A), the specifics of the 4D Cross-scale Embedding Layer (4D CEL) (Section II-B), and the 4D fMRI CrossFormer block (Section II-C). Further, we discuss the variants of the 4DfCF (Section II-D).

TABLE I
CONFIGURATION OF THE EMBEDDING LAYER FOR EACH STAGE IN THE 4D FMRI CROSSFORMER

Type	Kernel	Stride	Dim
3D Conv.	4x4x4	4x4x4	$\frac{D_t}{2}$
3D Conv.	8x8x8	4x4x4	$\frac{D_t}{4}$
3D Conv.	16x16x16	4x4x4	$\frac{D_t}{8}$
3D Conv.	32x32x32	4x4x4	$\frac{D_t}{16}$

A. 4D fMRI CrossFormer Overall Architecture

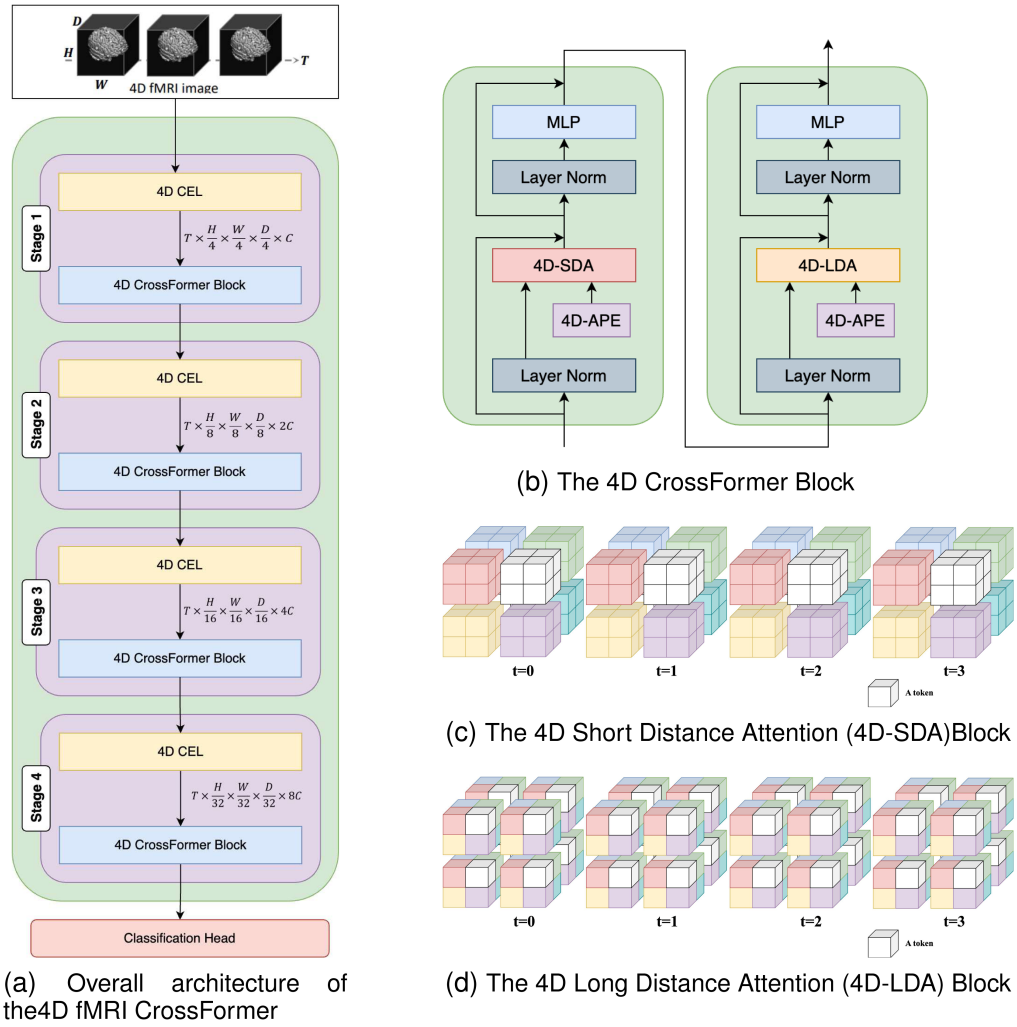
The 4DfCF is an advanced vision transformer designed specifically to process 4D fMRI images that incorporate both spatial and temporal dimensions. This model integrates the temporal dimension with three spatial dimensions to efficiently handle the complexity of 4D data. The architecture is structured hierarchically into four stages, each comprising a 4D CEL (Section II-B) and several 4D CrossFormer Blocks (Section II-C). A specialized head (i.e., the classification head in Fig. 1(a)) follows the final stage, which accounts for a specific classification task. The comprehensive structure of the 4D fMRI CrossFormer is shown in Fig. 1(a).

Fig. 1(b) provides an in-depth look at the structure of a 4D CrossFormer block, while Fig. 1(c) and (d) show the configurations of the 4D-SDA and 4D-LDA blocks, respectively. Details of these components are discussed in the following subsections.

B. 4D Cross-Scale Embedding Layer

The 4D CEL is a critical component of the 4DfCF, responsible for generating input tokens for each stage. This layer processes 4D fMRI images by sampling patches using kernels of various sizes. The stride of these kernels is maintained constant to ensure that the number of tokens generated remains consistent across different scales, as illustrated in Fig. 2 and Table I.

Taking the first 4D CEL as an example, ahead of *Stage-1*, it receives a 4D fMRI image as input. The input image $I \in \mathbb{R}^{T\times H\times W\times D}$, where T represents the temporal dimension and H, W, and D denote the spatial dimensions, is then sampled by four different kernels: $4\times4\times4$, $8\times8\times8$, $16\times16\times16$, and $32\times32\times32$, all with the same stride $4\times4\times4$. For each kernel size k, the embedding dimension d_k is allocated based on the computational budget. Larger kernels have fewer dimensions to control the overall computational cost. This allocation ensures



(a) Overall architecture of the4D fMRI CrossFormer

(b) The 4D CrossFormer Block

(c) The 4D Short Distance Attention (4D-SDA)Block

(d) The 4D Long Distance Attention (4D-LDA) Block

Fig. 1. The structure of the proposed 4D fMRI CrossFormer and its components.

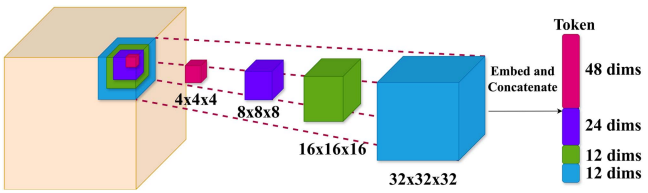


Fig. 2. Schematic representation of the 4D Cross-Scale Embedding layer (4D CEL).

that the tokens captured multi-scale features from the fMRI data, providing a rich representation. The dimensions are listed in Table I. The process of sampling and embedding can be mathematically described as:

- *Sampling Patches:* For each kernel size k, patches were sampled from the input image. Let P_k represent the patches sampled by a kernel of size $k \times k \times k$.
- *Embedding Patches:* Each patch P_k is then passed through a convolutional layer to generate embeddings E_k:

$$E_k = Conv_k(P_k) \tag{1}$$

- *Concatenating Embeddings:* The embeddings from different scales are concatenated to form a single token:

$$E = Concat(E_4, E_8, E_{16}, E_{32}) \tag{2}$$

where $E \in \mathbb{R}^{T \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times D_t}$ is the concatenated embedding with dimension D_t.

Fig. 2 illustrates the embedding and concatenation processes, where each of the four convolutional layers extracts patches of different sizes from the input image, and these patches are then embedded and concatenated to form comprehensive token representations. The 4D CELs in subsequent stages (*Stage-2/3/4*) operate similarly, but with different configurations to form a hierarchical structure. Specifically, the 4D CELs for *Stage-2/3/4* use two different kernels: $2 \times 2 \times 2$ and $4 \times 4 \times 4$, with a stride of $2 \times 2 \times 2$. This configuration reduces the number of embeddings to a quarter, ensuring a pyramid structure in which the spatial resolution decreases and the depth increases across the stages. Mathematically, $E^{(I)}$ denotes the embeddings at the i-th stage. The tokens for *Stage-2/3/4* are computed as:

$$E^{(i+1)} = Concat(Con_{2\times2\times2}(E^{(i)}), Conv_{4\times4\times4}(E^{(i)})) \tag{3}$$

where i represents the current stage, and $Conv_{k\times k\times k}$ denotes the convolutional operation with a kernel size $k \times k \times k$.

By progressively refining the feature representations at each stage, the 4D CELs enable the model to effectively capture both local and global dependencies in the 4D fMRI data. This hierarchical approach allows the model to learn intricate spatiotemporal patterns, enhancing its ability to analyze complex fMRI images [18].

C. 4D CrossFormer Block

The 4D CrossFormer Block is the core computational unit in the 4DfCF, which was designed to process 4D fMRI data by capturing both local and global attention. Each 4D CrossFormer Block consists of a 4D Short-Distance Attention (4D-SDA), a 4D Long-Distance Attention (4D-LDA), 4D Absolute Positional Embedding (4D-APE), and Multi-Layer Perceptron (MLP). As shown in Fig. 1(b), the 4D-SDA and 4D-LDA appear alternately in different blocks, and the 4D-APE module works in both 4D-SDA and 4D-LDA for obtaining embeddings' position representations. Following the previous vision transformers, residual connections were used in each block [19].

1) 4D Long and Short Distance Attention: The self-attention module in the 4D CrossFormer Block is split into two parts: 4D-SDA and 4D-LDA. This dual approach aids in the capture of both local and long-range dependencies within the 4D fMRI data. In 4D-SDA, adjacent embedding within $G \times G \times G \times G$ neighborhood are grouped together. For instance, in a 4D fMRI image with temporal dimension T and spatial dimensions H, W, and D, then G is the grouping size for the 4D-SDA. Fig. 1(c) provides an example where $G = 2$. The embeddings were grouped based on their proximity, ensuring that each group contained closely related data points. The 4D-LDA focuses on capturing long-range dependencies by attending to embeddings sampled at fixed intervals I. For example, Fig. 1(d) illustrates the 4D-LDA grouping, where $I = 2$. All embeddings with red borders belong to one group, whereas those with yellow borders belong to another group. Interval I determines the grouping size for 4D-LDA, which can be computed as $G = \frac{S}{I}$, where S is the dimension of the input. By grouping embeddings in this manner, both the 4D-SDA and 4D-LDA employ vanilla self-attention within each group. This significantly reduces the memory and computational costs of the self-attention module from $O(S^4)$ to $O(S^2\, G^2)$. It is important to highlight that the effectiveness of 4D-LDA was significantly enhanced by inputting cross-scale embeddings. For example, as illustrated in Fig. 1(d), the small-scale patches of two red embeddings were not adjacent. This non-adjacency makes it challenging to establish relationships without additional contextual information. In simpler terms, constructing dependencies between these embeddings based solely on small-scale patches (i.e., single-scale features) is difficult. Conversely, when large-scale patches are used, they can provide more ample context, making it easier to link distant embeddings. This approach simplifies the computation and enhances the meaningfulness of the long-distance cross-scale attention [18].

2) 4D Absolute Positional Embedding (4D-APE): In the 4DfCF, we employed an absolute positional embedding scheme rather than relative position biases to encode positional information. Although relative positional biases have been effective in previous models, the nature of 4D data makes absolute positional embeddings more computationally feasible and effective. Given the large scale of the 4D fMRI data, absolute positional embeddings provide clear advantages in terms of computational cost and performance [20]. For each stage of the 4DfCF, we added learnable positional embedding immediately after the patch-merging step. This approach ensures that both the spatial and temporal positional information are adequately represented. Following the method outlined in [21], we added positional embeddings separately for the spatial and temporal dimensions. Given an input tensor with dimensions $T \times H' \times W' \times D' \times C'$, we define the spatial and temporal positional embedding tensors as follows:

$$E_{spatial} \in \mathbb{R}^{1\times H'\times W'\times D'\times C'} \tag{4}$$

$$E_{temporal} \in \mathbb{R}^{T\times 1\times 1\times 1\times C'} \tag{5}$$

These embeddings are added to the input tensor I using broadcasting:

$$I' = I + E_{spatial} + E_{temporal} \tag{6}$$

where I' is the resulting tensor after adding the positional embeddings.

The absolute positional embedding function 4D-APE for a 4D tensor can be formulated as:

$$A_{i,j,k,t} = APE(i, j, k, t) \tag{7}$$

where $APE(i, j, k, t)$ generates positional embeddings based on the indices (i, j, k, t), which correspond to the temporal and spatial dimensions of the input tensor.

The inclusion of 4D-APE enhanced the attention mechanism of the 4D-LDA and 4D-SDA. The attention computations were as follows:

$$Attn = softmax\left(\frac{QK^T}{\sqrt{d}} + APE_{4D-LDA/SDA}\right)V \tag{8}$$

where Q, K, and V represent the *query, key, value* in the self-attention module, respectively. Further, where $\sqrt{d}$ is a constant normalizer, and $APE_{4D-LDA/SDA}$ is the 4D absolute positional embedding applied in both 4D-SDA and 4D-LDA.

D. Variants of 4D fMRI CrossFormer

Table II lists the detailed configurations of the two primary variants of the 4D fMRI CrossFormer model: tiny (4D fMRI CrossFormer-T) and base (4D fMRI CrossFormer-B). Both variants were structured hierarchically into four stages, each comprising cross-embedding layers and MLPs incorporating the 4D-SDA and 4D-LDA mechanisms. The main differences between these variants lie in their token dimensions, group sizes, and the number of layers used at each stage. The 4D fMRI CrossFormer-B features larger token dimensions and more layers at certain stages than the 4D fMRI CrossFormer-T, thereby enabling it to capture more detailed features and manage larger and more complex datasets. These differences allow the 4D

TABLE II
VARIANTS OF THE 4D FMRI CROSSFORMER. FOR SIMPLICITY, THE CLASSIFICATION HEAD (GLOBAL AVERAGE POOLING AND A LINEAR LAYER) IS OMITTED. D, G, AND I REPRESENT THE TOKEN DIMENSIONS, GROUP SIZES, AND INTERVALS FOR 4D-SDA AND 4D-LDA, RESPECTIVELY.

Stage	Layer	Tiny (-T)	Base (-B)
Stage-1	Cross Embed.	Kernel size: 4^3, 8^3, 16^3, 32^3, Stride = 4	
	4D-SDA/LDA	$\begin{pmatrix} D_1 = 64 \\ G_1 = 7 \\ I_1 = 8 \end{pmatrix} \times 1$	$\begin{pmatrix} D_1 = 96 \\ G_1 = 7 \\ I_1 = 8 \end{pmatrix} \times 2$
Stage-2	Cross Embed.	Kernel size: 2^3, 4^3, Stride = 2	
	4D-SDA/LDA	$\begin{pmatrix} D_2 = 128 \\ G_2 = 7 \\ I_2 = 4 \end{pmatrix} \times 1$	$\begin{pmatrix} D_2 = 192 \\ G_2 = 7 \\ I_2 = 4 \end{pmatrix} \times 2$
Stage-3	Cross Embed.	Kernel size: 2^3, 4^3, Stride = 2	
	4D-SDA/LDA	$\begin{pmatrix} D_3 = 256 \\ G_3 = 7 \\ I_3 = 2 \end{pmatrix} \times 8$	$\begin{pmatrix} D_3 = 384 \\ G_3 = 7 \\ I_3 = 2 \end{pmatrix} \times 18$
Stage-4	Cross Embed.	Kernel size: 2^3, 4^3, Stride = 2	
	4D-SDA/LDA	$\begin{pmatrix} D_4 = 512 \\ G_4 = 7 \\ I_4 = 1 \end{pmatrix} \times 6$	$\begin{pmatrix} D_4 = 768 \\ G_4 = 7 \\ I_4 = 1 \end{pmatrix} \times 2$

fMRI CrossFormer-B to balance the model complexity and computational efficiency effectively, ensuring robust performance in diverse tasks, such as disease classification and prediction.

III. EXPERIMENTS

A. Experiment Settings

1) Datasets: The performance of the proposed architecture was evaluated using three well-known benchmark datasets. (1) The ADHD-200 dataset is a multi-site collection aimed at understanding ADHD through brain imaging and related behavioral outcomes. This dataset is publicly available to the scientific community, but requires authorization for access [22]. After quality control, we used the fMRI data from five specific sites, including *NYU*, *OHSU*, *Peking 2*, and *NeuroIMAGE*. The data statistics after quality control are summarized in Table III. (2) The ADNI dataset is another significant dataset aimed for understanding Alzheimer's Disease (AD) using brain imaging data [23]. Specifically, we utilized fMRI imaging data from the *ADNI2* and *ADNI3* phases. These phases provide extensive longitudinal data essential for studying AD progression. The data statistics for the post-quality control are presented in Table III. (3) The ABIDE dataset focuses on Autism Spectrum Disorder (ASD) by aggregating brain imaging data from multiple sites [15]. This dataset helps to explore the neural basis of ASD through various neuroimaging modalities. For our analysis, we utilized data from ten specific sites: *CIT*, *MU*, *KKI*, *LMUM*, *NYU*, *UCLA*, *UUSM*, *SDSU*, *SU*, and *TCHS*. These sites provide resting-state fMRI data crucial for understanding connectivity patterns in individuals with ASD. The data statistics after quality control are presented in Table III.

In these datasets, $DX = 0$ indicates subjects not diagnosed with ADHD, AD, or ASD. Conversely, $DX = 1$ indicates subjects diagnosed with ADHD, AD, or ASD. The distribution of participants across different sites and diagnostic categories is presented in Table III. This distinction is critical for our binary classification task, in which the goal is to differentiate between typically developing individuals ($DX = 0$) and those with the target diseases ($DX = 1$).

TABLE III
DETAILED STATISTICS FOR THE FMRI DATASETS USED IN THE STUDY ACROSS DIFFERENT SITES AND DIAGNOSTIC CATEGORIES

Diseases	Datasets	DX=0	DX=1	Total
ADHD-200	NYU	98	118	216
	OHSU	42	37	79
	Peking 2	32	35	67
	NeuroIMAGE	23	25	48
	Total	**195**	**215**	**410**
AD	ADNI2	148	243	391
	ADNI3	139	149	288
	Total	**287**	**392**	**679**
ABIDE	CIT	19	19	38
	MU	55	55	110
	KKI	33	22	55
	LMUM	33	24	57
	NYU	105	79	184
	UCLA	33	49	82
	UUSM	43	58	101
	SDSU	22	14	36
	SU	20	20	40
	TCHS	25	24	49
	Total	**388**	**364**	**752**

2) Pre-Processing: For fMRI pre-processing, we employed the NIAK pipeline, a sophisticated and comprehensive tool designed for neuroimaging data analysis [24]. The NIAK includes several critical pre-processing steps to ensure that high-quality data are suitable for subsequent analyses. These steps encompass correction for slice-timing differences, correction for rigid body motion, correction for physiological noise, using methods such as CORSICA, and the high-pass filtering of slow-time drifts. Additionally, each dataset was aligned with a T1-weighted anatomical scan for enhanced structural accuracy, followed by a nonlinear transformation to the Montreal Neurological Institute (MNI) space using the CIVET pipeline, which includes surface-based cortical extraction and volumetric processing [25]. Functional volumes were subsequently resampled to a 3 mm isotropic resolution, and spatially smoothed using a 6 mm Gaussian kernel to minimize anatomical variability and enhance signal detection. Post-pre-processing steps involved low-pass filtering to remove high-frequency noise, head movement correction to account for subject motion during scans, and artifact removal by regressing signals from non-gray matter tissues using the aCompCor method [26]. This thorough pre-processing pipeline ensures that the data are of the highest quality for subsequent functional analyses and statistical modeling. These steps were implemented through the integration of the Pipeline System for Octave and Matlab (PSOM) and the Canadian Brain Research and Integrated Neuroimaging (CBRAIN), thus facilitating efficient high-throughput processing in a distributed computing environment. This approach not only maximizes computational efficiency, but also ensures the reproducibility and accessibility

of neuroimaging data processing across different computing infrastructures [27]. For each of the 4D fMRI volumes, we globally normalized the brain images over four dimensions, excluding the background regions. We then filled the background with the minimal voxel intensity value. To easily divide the volume into patches for the 4D fMRI CrossFormer, we changed the 3D volume to a shape of $96 \times 96 \times 96$ by cropping and padding on the background. To evaluate the performance of the ROI-based models, following the pre-processing steps of a previously established methodology, we applied the AAL atlas to each fMRI volume to obtain the time series data for each ROI [28]. Subsequently, we processed this ROI series to generate functional connectivity, which involved the computation of the Pearson's correlation coefficients to construct a correlation matrix. The correlation matrix was then Fisher-transformed, serving as the input for the ROI-based models. To ensure a robust evaluation, three random splits of the dataset were constructed, with a ratio of 70% for training, 15% for validation, and 15% for testing. We subsequently report the average performance across these three splits to account for the variability in each dataset. Due to the ADNI dataset lacking site-specific data and consisting only of ADNI2 and ADNI3 projects, as well as the fact that conducting experiments by site for the ADNI and ABIDE datasets would result in either unbalanced data (*e.g.*, ADNI2) or overly small datasets (*e.g.*, SDSU), we opted to merge these datasets for experiments. To better demonstrate the model's performance, the ADHD-200 dataset was evaluated by site, while the ADNI and ABIDE datasets were integrated for performance metrics evaluation in Section III-C4.

B. Implementation Details

We employed the 4DfCF model, specifically designed to handle the complexity of 4D fMRI data by effectively integrating temporal and spatial information.

1) Loss Function: A Binary Cross-Entropy (BCE) loss function was used for the binary classification task. The BCE loss is well suited for binary classification problems, for which the objective is to distinguish between two classes ($DX = 0$ for the normal case and $DX = 1$ for the diseased case). BCE loss is defined as follows:

$$\mathcal{L}^{i}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (9)$$

where N is the number of samples, y_i is the true label of the i-th sample, and p_i is the predicted probability that the i-th sample belongs to a positive class ($DX = 1$). This loss function penalizes the model more when it is confident of an incorrect prediction, thereby encouraging it to produce accurate probability estimates.

2) Training Procedure: The model was trained using the Adam optimizer, with an initial learning rate of 0.0001. The learning rate decayed by a factor of 0.5 every 20 epochs, to ensure that the model fine-tuned its weights effectively over time [29]. Early stopping based on the validation loss was employed to prevent overfitting, while training was halted if no improvement was observed over a predefined number of epochs. The batch size was set to 16, and the model was trained for a maximum of 100 epochs.

To handle memory constraints, dataset splitting was first performed at the subject level, ensuring that all subsequences from a subject were assigned exclusively to either the training, validation, or test set. After this split, each subject's entire fMRI volume was divided into 20 subsequences, which were treated as individual data points during training. For inference, the logits from all subsequences of a subject were averaged to yield a single subject-level prediction. For pre-training and transfer experiments (*e.g.*, ABIDE $\rightarrow$ ADNI), the datasets were strictly separated at the dataset level: no subjects or subsequences from the target dataset were exposed during pre-training. Fine-tuning and evaluation were then performed exclusively on the target dataset, ensuring that no information leakage occurred between pre-training and downstream testing.

3) Baselines: To evaluate the performance of the proposed 4DfCF, we compared it with four state-of-the-art baseline models: including pretrained models, CNN-based and transformer models. *Brain Language Model (BrainLM):* his model [30] leverages a transformer-based architecture trained on a large-scale fMRI corpus, using spatiotemporal masking for self-supervised pretraining, and is capable of predicting clinical variables and decoding functional networks with high accuracy. *Self-Supervised Learning for Brain Dynamics (SSL-CSM):* This framework [31] adopts causal language modeling to pretrain transformers on neuroimaging datasets spanning multiple acquisition sites and tasks, achieving superior generalization in mental state decoding tasks. *3D-CNNs Classifier:* The model, proposed in [32], leverages a 3D Convolutional Neural Network (CNN) architecture designed for classifying diseases using resting-state fMRI data. *Brain Network Transformer (BNT):* This model [33] uses a Transformer-based architecture for brain network analysis by treating the brain networks as graphs. *Transformer Framework for fMRI (TFF):* TFF [34] is a self-supervised transformer Framework that reconstructs 3D fMRI volume data during pre-training, and is optimized for specific tasks during fine-tuning.

SwiFT (Swim 4D fMRI Transformer): SwiFT [14] is a Swin Transformer variant extended to 4D, using local window attention to capture fine-grained spatiotemporal patterns. It employs multi-head self-attention and feedforward networks with global attention blocks to capture long-range dependencies. Further, SwiFT demonstrates robust performance in sex classification and age regression tasks on large-scale fMRI datasets, making it an advanced baseline for comparison.

These baseline methods provided robust points of comparison for evaluating the effectiveness of the proposed 4DfCF. By comparing these established models, we can demonstrate the advantages of our model for handling complex spatiotemporal patterns in fMRI data.

C. Experimental Results

This section provides a comprehensive evaluation of the proposed 4DfCF compared with state-of-the-art baseline models. First, to verify the impact of each module in our proposed model on overall performance, we conducted ablation experiments

TABLE IV
ABLATION EXPERIMENTS ON ADNI DATASETS TO EVALUATE THE DESIGN CHOICES OF THE 4DFCF

Method	ACC	AUC	Precision	Recall	F1-Score	Time/# Param.
4D fMRI CrossFormer	**95.75**	**96.24**	**95.06**	**97.59**	**96.25**	**15.2h/10.34M**
Stage-1	50.53	51.36	51.86	49.58	53.07	4.3h/1.21M
Stage-1∼2	64.72	65.27	64.81	60.50	61.89	6.0h/3.74M
Stage-1∼3	76.42	77.94	74.80	79.41	73.79	9.1h/8.13M
Stage-1∼5	**96.12**	96.01	**96.04**	96.62	96.22	23.6h/18.46M
with 4D-SDA & 4D-SDA	85.61	81.58	82.90	79.99	83.26	14.7h/10.01M
with 4D-LDA & 4D-LDA	82.51	82.36	80.48	78.75	81.20	15.0h/11.08M
with 4D-LDA & 4D-SDA	95.07	94.16	94.72	96.17	96.37	15.9h/10.57M
with 4D-CPE	86.47	85.58	90.44	87.99	89.41	20.2h/19.51M
with 4D-DPB	90.15	87.72	91.60	85.48	91.08	17.1h/12.68M
with 4D-RPE	91.53	91.86	90.51	92.68	92.83	19.4h/13.19M

using the ADHD-200 dataset. Subsequently, all experiments were conducted using three benchmark datasets: ADHD-200, ADNI, and ABIDE. The training accuracy and loss were assessed across multiple sites within each dataset to demonstrate the effectiveness and superiority of our model. In addition to comparing the training accuracy and loss, the model's performance in disease classification and diagnosis was evaluated using evaluation metrics. Furthermore, the training accuracy of the pre-trained model was compared with that of the non-pre-trained model. Further, the 4D fMRI CrossFormer-T variant was evaluated, focusing on the speed of convergence in terms of training accuracy and training loss, as well as the optimal results achieved, demonstrating its efficiency through a comparison of model training efficiency to further substantiate our findings. Finally, using an Integrated Gradient with Smoothgrad sQuare (IG-SQ) implemented in the Captum framework [16], [37], brain regions with high explanatory power for the classification task were identified. Detailed analysis is provided in the "Interpretation Results" section, where we present the 4D IG-SQ maps from test sets, highlight the spatial patterns across subjects, and identify key brain regions contributing to successful disease classification.

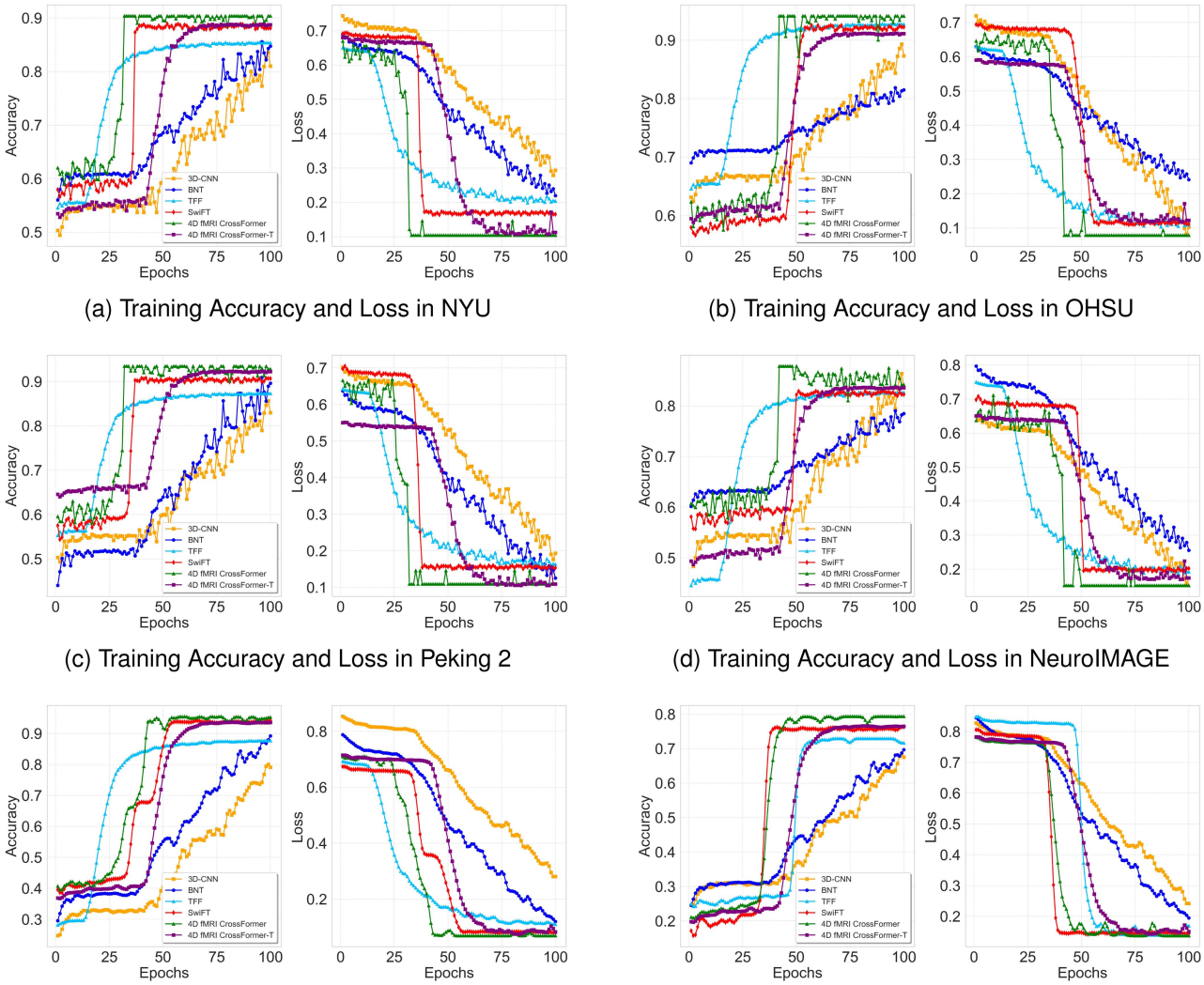
1) Ablation Experiment: The ablation experiments conducted on the ADNI dataset, as summarized in Table IV, were designed to evaluate the impact of different components and configurations of the proposed 4D fMRI CrossFormer model.

Effect of the number of stages: The results show that model performance improved steadily as the number of stages increased. Accuracy rose from 50.53% with Stage-1 to 64.72% with Stages 1–2 and further to 76.42% with Stages 1–3. The four-stage configuration achieved a substantial improvement, reaching 95.75% accuracy with a training time of 15.2 hours and 10.34 M parameters. Although the five-stage version slightly increased accuracy to 96.12% , it required far greater resources, taking 23.6 hours and 18.46 M parameters. These results indicate that the four-stage configuration offers the best trade-off between performance and efficiency.

Effect of attention ordering: The results show that repeating the same attention module is suboptimal: two consecutive 4D-SDA blocks achieved 85.61% accuracy, while two consecutive 4D-LDA blocks dropped to 82.51% . Combining 4D-SDA and 4D-LDA improved performance, but the order was critical. Applying 4D-LDA before 4D-SDA yielded 95.07% accuracy with 15.9 hours of training and 10.57 M parameters. By contrast, the reverse order, which corresponds to the final 4DfCF configuration, achieved the best overall performance of 95.75% accuracy with lower cost, requiring only 15.2 hours and 10.34 M parameters. These results confirm that the 4D-SDA → 4D-LDA sequence offers both high accuracy and computational efficiency, making it the preferred design.

Effect of positional encoding: The proposed 4D-APE outperformed alternative positional encoding methods. It achieved 95.75% accuracy with a training time of 15.2 hours and 10.34 M parameters. By comparison, 4D-Convolution-Base Positional Encoding (4D-CPE) [35] reached 86.47% accuracy with 20.2 hours of training and 19.51 M parameters, 4D-Dynamic Position Bias (4D-DPB) [17] achieved 90.15% accuracy with 17.1 hours and 12.68 M parameters, and 4D-Relative Positional Encoding (4D-RPE) [36] obtained 91.53% accuracy with 19.4 hours and 13.19 M parameters. These results demonstrate that 4D-APE provides superior accuracy while maintaining computational efficiency. In summary, the ablation experiments confirmed that the optimal configuration of the proposed 4D fMRI CrossFormer includes 4 stages, the combination of 4D-SDA and 4D-LDA (in sequence), and the 4D-APE for position embedding. This configuration achieves outstanding performance with efficient training time and parameter usage.

2) Training Accuracy and Training Loss: In our experiments, we compared the 4DfCF, including its tiny variant, with baseline models, such as 3D-CNNs, BNT, TFF, and SwiFT across various epochs. The results show that the 4DfCF consistently achieves a higher training accuracy, lower training loss, and reaches the optimum faster than the baselines. This was evident in all three datasets: the ADHD-200, ADNI, and ABIDE. For the ADHD-200 dataset, we conducted experiments at four specific sites: NYU, OHSU, Peking 2, and NeuroIMAGE. Fig. 3(a) to (d) show the training accuracy and losses for these sites. Our 4DfCF demonstrated superior performance at the NYU site, achieving higher accuracy and lower loss throughout the training epochs compared to the 3D-CNNs, BNT, and SwiFT. This improvement was especially noticeable after 20 epochs, where the accuracy of 4DfCF continued to increase steadily, whereas that of the other models plateaued. Similar trends were observed for the OHSU dataset, where 4DfCF significantly outperformed the baselines, maintaining a more stable and rapid decrease in loss, with a consistent increase in accuracy. In the Peking 2 dataset, our model exhibited a faster convergence and reached higher accuracy levels earlier during the training process. The final accuracy of the 4DfCF was significantly higher than that of the other models. For NeuroIMAGE, the 4DfCF again showed a robust performance, achieving the lowest loss and highest accuracy among all models, particularly excelling in the later epochs. In particular, the 4D fMRI CrossFormer-T variance, demonstrated efficient convergence, reaching optimal performance quicker than the SwiFT, highlighting its effectiveness in managing the complexity of fMRI data. Overall, the 4DfCF's ability to manage and learn from the ADHD-200



(a) Training Accuracy and Loss in NYU

(b) Training Accuracy and Loss in OHSU

(c) Training Accuracy and Loss in Peking 2

(d) Training Accuracy and Loss in NeuroIMAGE

(e) Training Accuracy and Loss in ADNI

(f) Training Accuracy and Loss in ABIDE

Fig. 3. Comparison of the Training Accuracy and Loss on the ADHD, ADNI, and ABIDE Datasets. The x-axis represents the epoch. The y-axis of the left sub-figure in each figure represents the training accuracy, and the y-axis of the right sub-figure in each figure represents the training loss.

dataset's complex and high-dimensional data is evident, providing significant performance advantages over the baseline models.

For the ADNI dataset, we evaluated the models using fMRI data from the ADNI2 and ADNI3 phases. Fig. 3(e) presents the training accuracy and loss. Overall, the 4DfCF achieved higher accuracy and lower loss throughout the training process compared to the baseline models. The accuracy improvement of the model was more pronounced after 30 epochs, indicating its superior capability at extracting relevant features from fMRI data. The loss metric consistently decreased at a faster rate, demonstrating the efficiency of our model in learning the complex patterns associated with Alzheimer's Disease. Notably, the 4D fMRI CrossFormer-T variant also showed faster convergence than the SwiFT, further validating the robustness of the model design. For the ABIDE dataset, Fig. 3(f) presents a comparison of the training accuracy and loss. The 4DfCF outperformed the baseline models across multiple sites within the ABIDE dataset. The model exhibited rapid convergence, reaching higher accuracy levels much earlier in the training process. The final training accuracy was significantly higher, and the loss was lower than that of 3D-CNNs, BNT, and SwiFT, underscoring the model's robustness and efficiency in handling both diverse and high-dimensional fMRI datasets. In particular, the 4D fMRI CrossFormer-T variant, demonstrated significant improvements in convergence speed, achieving better performance metrics earlier than the SwiFT. Overall, these results validate the effectiveness of the proposed method in enhancing the analysis and understanding of complex brain disorders using fMRI data. The 4DfCF, including its tiny variant, not only achieved superior accuracy and loss metrics, but also exhibited faster and more stable convergence, particularly in comparison to the SwiFT model, highlighting the robustness and efficiency of the 4DfCF in managing high-dimensional fMRI data across various classification tasks.

3) Validation Performance: To evaluate the training stability and generalization capability of our proposed 4DfCF, we analyzed its training and validation performance across ADHD-200,

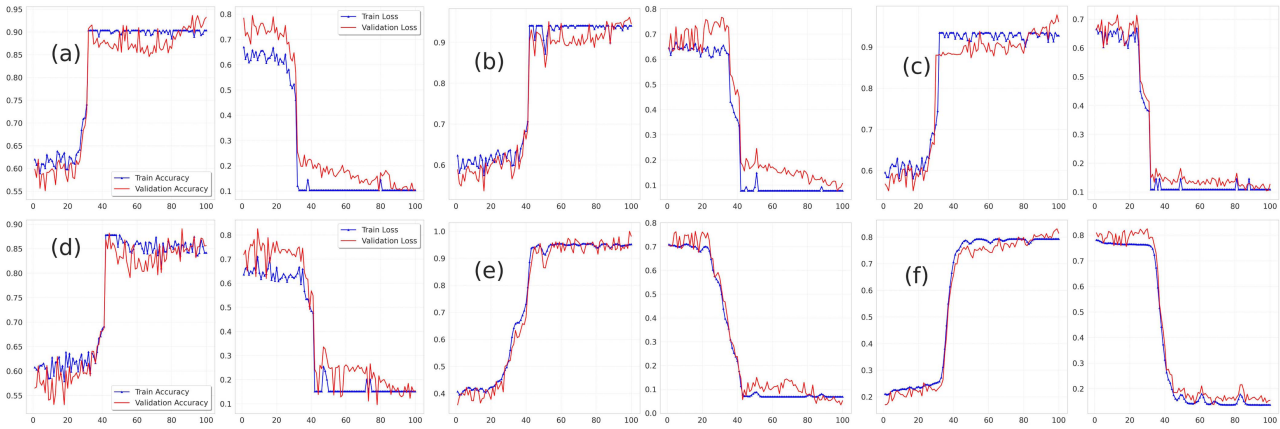


Fig. 4. Comparison of the Validation Curves on the ADHD, ADNI, and ABIDE Datasets. In each sub-figure (i.e., (a)–(f)), the x-axis represents the epoch, and the y-axis represents validation accuracy (left) or validation loss (right). Fig. 4 displays validation curves for six datasets, i.e., (a) ADHD/NYU, (b) ADHD/OHSU, (c) ADHD/Peking 2, (d) ADHD/NeuroIMAGE, (e) ADNI, and (f) ABIDE.

ADNI, and ABIDE datasets (Fig. 4). Overall, 4DfCF demonstrated rapid convergence with stable improvements in training accuracy and consistent reductions in training loss throughout epochs. Notably, validation accuracy closely followed training trends, steadily increasing while maintaining a low validation loss, indicating robust generalization without significant overfitting. These results illustrate the effectiveness of the 4DfCF in modeling complex spatiotemporal fMRI data and its reliability in clinical prediction tasks across diverse datasets.

4) Performance Metrics: We evaluated the effectiveness of our proposed 4DfCF using various performance metrics (ACC, AUC, precision, recall, and F1-Score) across different datasets, and compared them with baseline models (i.e., BrainLM, SSL-CSM, 3D-CNNs, BNT, TFF, and SwiFT). The results, summarized in Table. V, showed that the 4D fMRI CrossFormer consistently outperformed these baselines in all metrics across the ADHD-200, ADNI, and ABIDE datasets. Both the main 4DfCF and its smaller variant (4D fMRI CrossFormer-T) demonstrated superior performance. The models showed higher accuracy, better precision, improved recall, and better overall F1-Scores compared to the baseline models. Because all performance metrics were consistent, we used the F1-Score to compare the models. For example, on the ADHD-200 dataset, the main 4DfCF achieved an F1-Score of 91.04% , whereas the tiny variant achieved 89.52% , indicating that the tiny variant maintained a high level of performance despite its smaller size. On the ADNI dataset, the main 4DfCF achieved an F1-Score of 96.28% , compared to 93.54% for the tiny variant. This gap of approximately 2.74% further demonstrates the efficiency of the tiny variant in retaining most of its accuracy with fewer resources. In the ABIDE dataset, the main model achieved an F1-Score of 79.78% , while the tiny variant achieved 76.06% , representing a gap of approximately 3.72% , demonstrating the robustness of this tiny variant in high-dimensional fMRI data. These results highlight that while the main 4DfCF consistently delivers superior performance, the tiny variant achieves comparable results with only a minor reduction in accuracy. This makes the tiny variant an efficient alternative for scenarios in which computational resources are limited.

5) Model Efficiency: The efficiency of the proposed 4DfCF was assessed by comparing the number of parameters [38], Floating-point Operations Per second (FLOPs) [39], and throughput [40] with baseline models. The results, summarized in Table VI, demonstrate the efficiency of the 4DfCF. The 4DfCF has 10.34 M parameters, 3.45 G FLOPs, and a throughput of 98.61, making it significantly more efficient than the baseline models Brain-LM, SSL-CSM, 3D-CNNs, BNT, and TFF, which have much more parameters and FLOPs, resulting in a lower throughput, as shown in Table VI. The 4D fMRI CrossFormer-T variant further improved the efficiency with only 4.18 M parameters, 2.01 G FLOPs, and a throughput of 197.34. Although 4D fMRI CrossFormer-T may not achieve the same level of performance metrics (such as accuracy and recall) as SwiFT and 4DfCF, it excels in terms of resource utilization. For example, in the ADHD-200/NYU dataset, the F1-Score of the 4D fMRI CrossFormer-T was 89.52% , compared to 94.65% for the main 4DfCF, representing a performance loss of approximately 5.13% . Compared with the SwiFT, which had an F1-Score of 92.79% , the performance loss was actually a gain of approximately 4.27% . On the ADNI dataset, the F1-Score of the 4D fMRI CrossFormer-T was 93.54% , compared with 96.28% for the main 4DfCF, indicating a performance loss of approximately 2.74% . Compared with the SwiFT, which had an F1-Score of 90.23% , the performance loss showed a gain of approximately 3.31% . On the ABIDE dataset, the F1-Score of the 4D fMRI CrossFormer-T was 76.06% , compared to 79.78% for the main 4DfCF, representing a performance loss of approximately 3.72% . Compared with the SwiFT, which had an F1-Score of 79.34% , the performance loss was approximately 0.44% . These numbers demonstrate that while the 4D fMRI CrossFormer-T uses fewer computational resources and has faster training times than both SwiFT or the main 4DfCF model, it maintains a high level of performance, with only a minor reduction in accuracy. According to the results in Table V, 4D fMRI CrossFormer-T achieves comparable performance and, in some cases, even outperforms SwiFT, which is considered a high-level performance model. These efficiency metrics highlight that both the main and the tiny variants of the 4D fMRI CrossFormer

TABLE V
COMPARISON OF THE PERFORMANCE METRICS COMPARISON OF DIFFERENT METHODS AND DATASETS, HIGHLIGHTING THE ACCURACY, AUC, PRECISION, RECALL, AND F1-SCORE ACROSS ADHD-200, ADNI, AND ABIDE DATASETS

Dataset	Method	ACC	AUC	Precision	Recall	F1-Score
ADHD /NYU	BrainLM	76.62	75.23	72.52	77.35	77.52
	SSL-CSM	81.32	80.53	81.69	85.83	83.88
	3D-CNNs	83.55	83.39	85.21	82.53	83.85
	BNT	85.57	87.57	83.61	93.29	88.19
	TFF	85.45	79.57	78.03	81.35	79.66
	SwiFT	88.90	84.55	81.47	86.61	83.96
	4D fMRI CrossFormer-T	88.81	82.36	80.98	82.67	81.81
	4D fMRI CrossFormer	**90.40**	**90.49**	**85.51**	**97.32**	**91.04**
ADHD /OHSU	BrainLM	80.62	78.63	82.86	78.72	79.37
	SSL-CSM	82.41	81.66	79.42	84.25	84.63
	3D-CNNs	89.31	87.23	82.50	92.37	87.16
	BNT	81.86	87.98	87.52	88.31	87.91
	TFF	92.80	**92.75**	93.14	92.56	**92.85**
	SwiFT	92.56	87.30	86.29	87.82	87.05
	4D fMRI CrossFormer-T	91.18	92.47	86.68	**97.71**	91.86
	4D fMRI CrossFormer	**94.15**	91.77	**94.96**	88.67	91.71
ADHD /Peking 2	BrainLM	72.65	73.73	71.84	77.48	76.11
	SSL-CSM	86.21	83.77	81.86	86.42	83.46
	3D-CNNs	85.65	81.47	80.53	82.47	81.49
	BNT	90.93	85.82	84.23	86.19	85.20
	TFF	87.39	84.00	82.27	85.47	83.84
	SwiFT	90.80	92.40	91.43	94.19	92.79
	4D fMRI CrossFormer-T	92.29	89.38	85.07	94.45	89.52
	4D fMRI CrossFormer	**93.49**	**94.76**	**94.17**	**95.14**	**94.65**
ADHD /Neuro IMAGE	BrainLM	73.27	75.14	71.45	77.17	73.46
	SSL-CSM	80.34	79.83	78.80	81.49	79.92
	3D-CNNs	86.38	88.63	91.33	**85.96**	88.56
	BNT	79.01	79.45	74.56	83.97	78.99
	TFF	82.72	82.95	85.41	77.82	81.44
	SwiFT	82.90	82.61	81.69	82.80	82.24
	4D fMRI CrossFormer-T	83.64	88.16	91.37	84.65	87.88
	4D fMRI CrossFormer	**87.83**	**89.20**	**92.04**	85.34	**88.57**
ADNI	BrainLM	76.03	72.69	73.98	76.91	72.58
	SSL-CSM	79.07	80.24	76.95	80.31	81.68
	3D-CNNs	81.97	87.69	87.53	88.69	88.10
	BNT	91.58	89.13	90.09	87.84	88.86
	TFF	87.81	90.38	87.45	93.07	90.15
	SwiFT	94.47	89.98	92.51	88.15	90.23
	4D fMRI CrossFormer-T	93.66	93.58	93.02	94.08	93.54
	4D fMRI CrossFormer	**95.75**	**96.24**	**95.06**	**97.59**	**96.28**
ABIDE	BrainLM	60.42	58.13	61.42	60.90	59.36
	SSL-CSM	62.21	63.14	67.16	62.81	64.38
	3D-CNNs	69.78	63.61	64.09	62.45	63.26
	BNT	71.50	76.44	75.33	79.31	77.27
	TFF	72.97	72.66	79.08	66.99	72.53
	SwiFT	76.48	**79.32**	80.78	77.95	79.34
	4D fMRI CrossFormer-T	76.61	74.84	76.02	76.11	76.06
	4D fMRI CrossFormer	**79.37**	79.17	**80.17**	**79.49**	**79.78**

have highly effective computational resources and processing speed, making them suitable for large-scale fMRI data analysis.

6) Pre-Training Effects: Training a model from scratch randomly initializes its parameters. However, fine-tuning a pre-trained model initializes the training parameters in a good state, which can aid in the optimization process [41]. Herein, we examined the impact of pre-training on the performance of the proposed 4DfCF by comparing models pre-trained on one dataset, and fine-tuned on another, using models trained from scratch. The results depicted in Fig. 5 demonstrate the benefits of pre-training across different classification tasks. For example, pre-training on the ADHD-200 and ABIDE datasets before fine-tuning on the ADNI data (Fig. 5(b) and (e)) revealed a noticeable improvement in training accuracy compared to training from scratch. Specifically, pre-training on the ABIDE improved the training accuracy on the ADHD-200 and ADNI by 0.75% and 0.7% , respectively. Pre-training on ADNI resulted in improvements of 0.52% and 0.53% in the training accuracy for the ADHD-200 and ABIDE datasets, respectively. Similarly, models pre-trained on the ADHD-200 and then fine-tuned on the ABIDE data and ADNI(Fig. 5(a) and (d)) exhibited improved training accuracy, achieving an increase of 0.49% and 0.45% , respectively. The pre-trained models achieved higher accuracy earlier in the training process, indicating better convergence and faster learning.

TABLE VI
EFFICIENCY METRICS OF THE 4D FMRI CROSSFORMER COMPARED TO THE BASELINE MODELS: PARAMETERS, FLOPS, AND THROUGHPUT

Method	# Param.	FLOPs	Throughput
BrainLM	111M	7.34G	35.52
SSL-CSM	141M	16.31G	40.19
3D-CNNs	213M	10.65G	42.65
BNT	353M	21.53G	61.44
TFF	722M	41.54G	57.3
SwiFT	4.65M	2.62G	103.52
4D fMRI CrossFormer	10.34M	3.45G	98.61
4D fMRI CrossFormer-T	**4.18M**	**2.01G**	**197.34**

A detailed comparison showed that models pre-trained on the ABIDE dataset had better results, including a 0.72% and 0.7% improvement in training accuracy and faster convergence across all datasets. This superior performance is likely due to the larger dataset size of the ABIDE compared to the ADNI, and the significantly larger dataset size of the ADHD-200, which allowed the models to learn more features from the brain fMRI images during pre-training. Moreover, we observed that pre-training on ABIDE led to larger improvements when fine-tuned on ADHD-200 than on ADNI, and similarly, pre-training on ADHD-200 benefited ABIDE more than ADNI. This effect may be partially explained by the biological and demographic characteristics of the datasets: ABIDE and ADHD-200 involve younger populations with neurodevelopmental disorders, whereas ADNI mainly contains older adults with neurodegenerative conditions. Such differences in age and disease progression likely shape distinct fMRI patterns, which may partially explain why certain cross-dataset transfers generalize better than others.

These results highlight the advantages of using pre-trained models for fMRI data analysis. Pre-training on one dataset provides a solid foundation for fine-tuning on another dataset, leading to faster convergence, higher accuracy, and improved overall performance across various classification tasks.

7) Interpretation of the Results: Using an Integrated Gradient with Smoothgrad sQuare (IG-SQ) implemented in the Captum framework [16], [37], we identified brain regions showing high explanatory power for the classification task. The IG-SQ method combines the Integrated Gradients (IG) technique with

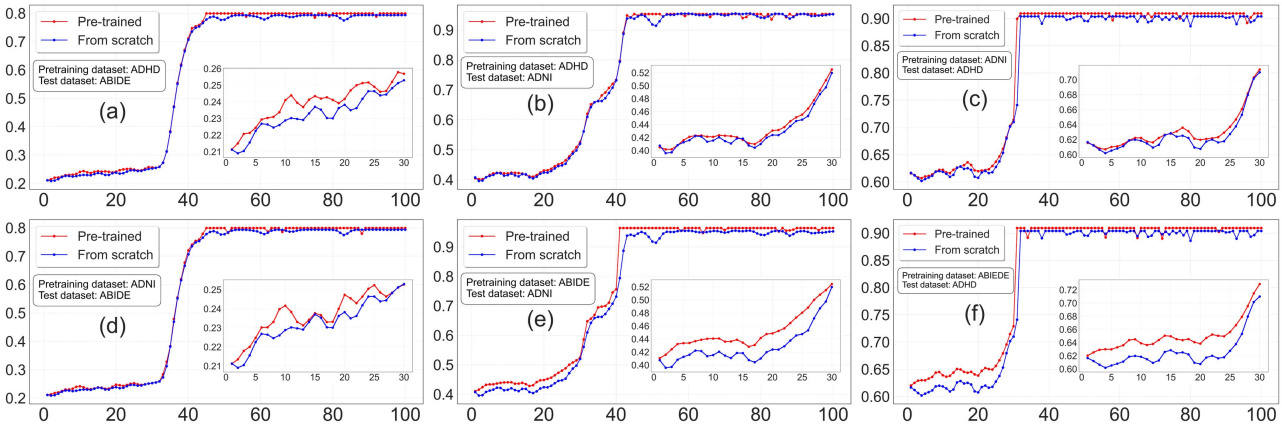


Fig. 5. Comparison of the Training Accuracy for Pre-trained Models, including evaluation of the impact of pre-training on ADHD, ADNI, and ABIDE datasets for improved convergence and performance. The x-axis represents the epoch, and the y-axis represents the training accuracy. Fig. 5(a) is the effect of ADHD-200 pre-trained on the ABIDE classification tasks, Fig. 5(b) is the effect of ADHD-200 pre-trained on the ADNI classification tasks, Fig. 5(c) is the effect of ADNI pre-trained on the ADHD-200 classification tasks, Fig. 5(d) is the effect of ADNI pre-trained on the ABIDE classification tasks, Fig. 5(e) is the effect of ABIDE pre-trained on the ADNI classification tasks, Fig. 5(f) is the effect of ABIDE pre-trained on the ADHD-200 classification tasks.

the Smoothgrad sQuare method to enhance interpretability and reduce noise in attribution maps. Specifically, IG computes the path integral of gradients by accumulating changes in the output with respect to scaled inputs along a straight path from a baseline input x_0 to the input x:

$$IG_i(x) = (x_i - x_{0i}) \int_{\alpha=0}^{1} \frac{\partial F(x_0 + \alpha \cdot (x - x_0))}{\partial x_i} \, d\alpha \quad (10)$$

where $F(x)$ is the model's output for input x, x_0 is the baseline input (*e.g.*, zero or mean values), and α is the scaling factor. Smoothgrad sQuare further enhances the IG maps by averaging attributions over multiple noisy variations of the input, improving focus on important features while suppressing irrelevant noise:

$$IG\text{-}SQ_i(x) = \frac{1}{n} \sum_{k=1}^{n} IG_i(x + \epsilon_k)^2 \quad (11)$$

where n is the number of noisy samples, and ϵ_k represents noise sampled from a Gaussian distribution.

We generated 4D IG-SQ maps from the test sets and filtered out the incorrectly predicted samples. To identify spatial patterns of the brain showing explanatory power across subjects, we normalized the IG-SQ maps and smoothed them with a Gaussian filter. Subsequently, we averaged the maps across subjects and over the time dimension. To further assess the robustness of these attribution patterns, we also conducted multiple experiments under different random seeds and across different subsets of subjects. The overall patterns were qualitatively consistent, with key regions repeatedly highlighted, although slight variations in intensity were observed. For illustration, we selected three representative subjects to display in Fig. 6, which exemplify the reproducible attribution maps observed across runs. Fig. 6 also illustrates the brain regions contributing to successful disease classification. For better comparative validation of our method's superiority, we selected a traditional CNN–based 3D-CNNs approach and a Vision Transformer–based SwiFT method as representatives. These two approaches underwent the same XAI experiments as 4DfCF, with results shown in Fig. 6.

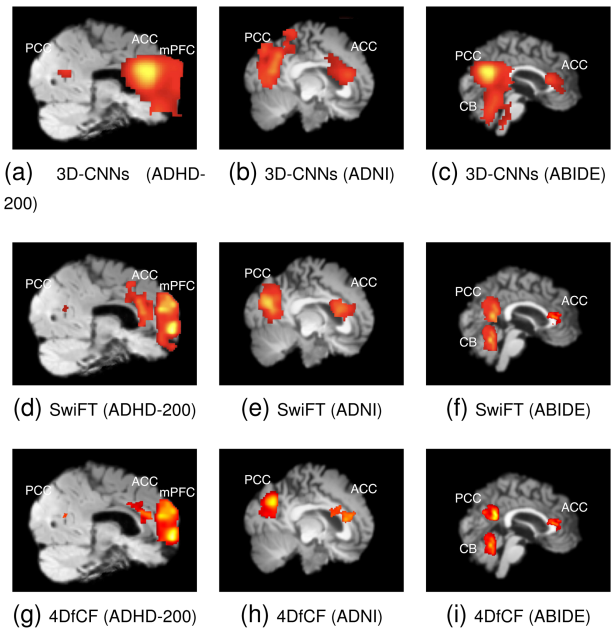


Fig. 6. Interpretation maps with Integrated Gradients (IG) for disease classification using 3D-CNNs, SwiFT and 4DfCF. (Sagittal plane).

In the ADHD dataset (Fig. 6(a), (d), and (g)), the medial PreFrontal Cortex (mPFC), Posterior Cingulate Cortex (PCC), and Anterior Cingulate Cortex (ACC) were identified as key regions. These nodes are critically involved in executive function, attention regulation, and default mode network (DMN) activity in ADHD [32]. Dysregulation of the mPFC contributes to impulsivity and attentional lapses [42], while aberrant ACC connectivity impairs conflict monitoring [43]. PCC abnormalities underpin mindwandering and reduced cognitive control commonly observed in ADHD patients [44]. When comparing attribution maps across the three methods, the 3DCNN

highlighted broad frontal and parietal areas without precise delineation of cingulate structures; SwiFT showed moderate focus on DMN regions but diffused attributions into adjacent white matter; in contrast, 4DfCF produced highly focal, anatomically coherent attributions in mPFC, ACC, and PCC, demonstrating superior sensitivity to ADHDrelated network dysfunctions.

In the ADNI dataset (Fig. 6(b), (e), and (h)), focusing on AD, the Posterior Cingulate Cortex (PCC) and Anterior Cingulate Cortex (ACC) were highlighted. The PCC is a core hub of the DMN and one of the earliest regions exhibiting hypometabolism and disrupted connectivity in AD [45]. ACC dysfunction correlates with deficits in cognitive control and decisionmaking in AD patients [46]. In XAI comparisons, 3DCNN attributions were diffuse across cortical and subcortical regions; SwiFT attributions concentrated on DMN nodes but lacked anatomical precision; by contrast, 4DfCF maps precisely delineated PCC and ACC, underscoring its efficacy in capturing clinically relevant degeneration patterns.

In the ABIDE dataset (Fig. 6(c), (f), and (i)), which investigates ASD, unique regions such as the Cerebellum (CB), PCC, and ACC were identified. Cerebellar abnormalities, particularly in Crus I/II, have been linked to motor coordination deficits and socialcognitive impairments in ASD [47], and disruptions in PCC and ACC connectivity relate to impaired selfreferential processing and social cognition [32]. Under the same XAI protocol, the 3DCNN and SwiFT models failed to resolve these finegrained features—3DCNN attributions were overly broad, and SwiFT diffused attention beyond key loci—whereas 4DfCF produced localized attributions in Crus I/II, PCC, and ACC, reflecting enhanced capability to detect ASDspecific anatomical signatures.

These results confirm the regions implicated in previous studies investigating the differences in diseased brains [32], [42], [43], [44], [45], [46], [47]. The highlighted regions in each dataset correspond to neural circuits most affected by the respective disorders, further validating the model's capability to capture meaningful and biologically relevant patterns.

IV. Discussion

Our experimental results clearly demonstrate that the proposed 4DfCF not only achieves superior quantitative performance—exhibiting higher accuracy, precision, recall, and F1-scores on the ADNI, ADHD-200, and ABIDE datasets—but also introduces significant architectural innovations. By leveraging cross-scale embeddings and a hierarchical attention structure, 4DfCF effectively integrates local fine-grained features with global contextual information, overcoming limitations seen in models like SwiFT that rely on localized attention mechanisms. Furthermore, the model exhibits faster and more stable convergence during training and offers an excellent accuracy–efficiency trade-off. Even its lightweight variant (4DfCF-T) maintains competitive performance with far fewer parameters, making it a versatile tool not only for achieving state-of-the-art results in disease classification but also for scalable neuroimaging analysis.

A. Clinical Impact

The strong performance of 4DfCF on neurological disorder classification tasks highlights its potential impact on clinical diagnosis of ADHD, Alzheimer's Disease, and Autism. By achieving high accuracy and F1-scores on fMRI data from patients, the model sets a new benchmark for computer-aided diagnosis. For instance, on the ADNI dataset for Alzheimer's, 4DfCF's classification accuracy reached about 95–96% , exceeding that of previous models. Such improvements in precision and recall translate to a lower risk of false negatives or false positives, which is crucial in a clinical setting. In practical terms, a model with this level of reliability could serve as a supportive diagnostic tool – for example, helping clinicians identify ADHD from resting-state fMRI with greater confidence or detect early Alzheimer's-related changes in brain function that might be missed by conventional analyses. This is particularly valuable for conditions like ADHD and ASD, where diagnoses currently rely heavily on subjective behavioral assessments. An objective, imaging-based evaluation from 4DfCF can provide additional evidence, potentially leading to earlier and more accurate diagnoses and informing treatment decisions in a clinical workflow.

Beyond improving diagnostic accuracy, 4DfCF also offers insights that can enhance neuroscience research and our understanding of brain disorders. Using interpretability techniques, we examined which brain regions the model considered most important for each condition. The results were encouraging: 4DfCF consistently highlighted regions that align with known neuropathology of the disorders. In ADHD, for example, the model identified the medial prefrontal cortex and the anterior/posterior cingulate cortices as key regions – areas involved in executive function and the default mode network, which are known to be dysregulated in ADHD. In Alzheimer's disease, 4DfCF focused on hubs of the default mode network such as the posterior cingulate cortex and adjacent precuneus, as well as the anterior cingulate. These regions are among the first to show functional decline in Alzheimer's, matching clinical observations of disrupted connectivity and metabolism in those areas. For ASD, the model brought attention to regions like the cerebellum and cingulate cortex, which have been implicated in motor-social processing and self-referential thought, respectively – both relevant to autism's core symptoms. These model-identified brain regions correspond closely with findings from prior neuroimaging studies, underscoring that 4DfCF is capturing biologically meaningful patterns rather than arbitrary features. This alignment builds trust in the model's decisions and demonstrates its utility in revealing disease biomarkers. From a research perspective, the ability to efficiently analyze whole-brain fMRI data means scientists can apply 4DfCF to large datasets to uncover subtle brain-behavior relationships. It can accelerate discovery of functional network differences in disorders, guide hypothesis generation about disease mechanisms, and ultimately contribute to precision neuroscience by linking specific brain dynamics to clinical outcomes.

B. Technical Impact

From a technical standpoint, 4DfCF offers significant advantages in computational efficiency, scalability, and adaptability,

marking a clear improvement over previous models. The model is highly lightweight relative to many earlier deep learning approaches for fMRI. It achieves top performance with on the order of only 10 million trainable parameters, whereas prior deep networks often used hundreds of millions. For instance, one baseline 3D-CNN model required over 200 M parameters and a recent transformer-based model (TFF) used 722 M, in contrast to 4DfCF's 10 M. This drastic reduction in model size leads to lower memory usage and faster computation. Empirically, our model's computational footprint is very low: it requires about 3.45 GFLOPs and achieves a throughput of 98 fMRI volumes per second in the standard configuration. The tiny variant (4DfCF-T) is even more efficient, with only 4 million parameters and a throughput near 197 volumes per second. In practical terms, such efficiency means that training and inference can be done on standard hardware and can scale to large datasets or high-throughput processing environments. Researchers and clinicians could deploy 4DfCF on hospital servers or research clusters and analyze data from many subjects in parallel, something that would be challenging with heavier models. These efficiency metrics confirm 4DfCF's suitability for large-scale neuroimaging analysis. As neuroimaging datasets continue to grow (*e.g.*, initiatives collecting thousands of fMRI scans), our model is well-equipped to handle the increase in data without a proportional increase in computation time, ensuring scalability for future applications.

Another key technical strength of 4DfCF is its adaptability and generalizability to different neuroimaging tasks. We demonstrated that the model can be pre-trained on one large fMRI dataset and then fine-tuned on another, leveraging shared patterns in brain activity to improve performance – a form of transfer learning. Notably, models pre-trained on one disorder (*e.g.*, in ABIDE) and then fine-tuned on another (*e.g.*, ADHD) converged faster and achieved higher accuracy than training from scratch. This indicates that 4DfCF can serve as a foundational model for fMRI analysis, carrying learned knowledge from one dataset to another. Indeed, our results highlight the potential of building large-scale foundation models for fMRI using 4DfCF's architecture, akin to how large pre-trained models are used in computer vision or NLP. Such a model could be adapted with minimal effort to new brain disorders, different age groups, or even other functional neuroimaging modalities, making it a versatile tool for the neuroscience community.

In addition, we considered the impact of acquisition variability across multisite datasets, such as differences in spatial resolution (*e.g.*, $2\times 2\times 2\,\mathrm{mm}^3$ *vs.* $3\times 3\times 3\,\mathrm{mm}^3$), repetition time (TR), and signal-to-noise ratio (SNR). Our standardized preprocessing pipeline—including nonlinear registration to MNI space, resampling all volumes to 3 mm isotropic resolution, slice-timing correction, temporal filtering, and ICA/aCompCor-based denoising—substantially reduced site-specific variability and brought the data into a comparable space. These procedures mitigate, though cannot fully eliminate, intrinsic disparities across acquisition protocols: resampling may lead to partial loss of fine-grained information, and TR differences cannot be completely corrected by interpolation. Nevertheless, the multi-scale embedding and hierarchical attention design of 4DfCF demonstrated robustness in this setting by effectively integrating both fine-grained and global spatiotemporal features. Future work could incorporate advanced harmonization approaches (*e.g.*, ComBat or domain adaptation) to further reduce residual site effects and strengthen generalizability in large-scale, real-world multi-center applications.

In summary, the technical innovations of 4DfCF in efficiency, scalability, robustness, and flexibility mark a substantial step forward over previous neuroimaging models, enabling more widespread and impactful use of deep learning in brain disorder analysis.

V. Conclusion

Investigating the spatiotemporal dynamics of the human brain is challenging due to the complexity of brain networks and the limitations of current analytical methods. In the present study, we introduced the 4D fMRI CrossFormer (4DfCF), a novel vision transformer designed to process high-dimensional 4D fMRI data. This model effectively integrates temporal and spatial dimensions, enabling it to handle intricate brain dynamics and predict cognitive or clinical outcomes. We tested the 4DfCF on three benchmark datasets: the ADHD-200, ADNI, and ABIDE. The results showed that our model consistently outperformed state-of-the-art baseline models in terms of accuracy, precision, recall, and F1-score. The 4D fMRI CrossFormer-T variant further demonstrated improved efficiency, used fewer resources, and achieved faster training times than the SwiFT and our main 4DfCF model. Pre-training experiments revealed that models pre-trained on one dataset and fine-tuned on another achieved faster convergence and higher accuracy. Notably, the models pre-trained on the ABIDE dataset performed best, likely owing to the larger dataset size, allowing more comprehensive feature learning. Efficiency metrics confirmed the robustness of the 4DfCF, indicating that it is highly suitable for large-scale fMRI data analysis. The superior performance, resource efficiency, and scalability of this model highlight its potential for advancing precision neuroscience. Future work will explore pre-training the 4DfCF on larger datasets to enhance its predictive capabilities. Finally, we identified brain regions showing high explanatory power in disease diagnosis tasks using an Integrated Gradient with Smoothgrad sQuare (IG-SQ) implemented in the Captum framework. These results demonstrate the critical value of the 4DfCF in elucidating the relationship between brain activity and disease diagnosis.

References

[1] Z. Liu et al., "Deep learning based brain tumor segmentation: A survey," *Complex Intell. Syst.*, vol. 9, no. 1, pp. 1001–1026, 2023.
[2] Y. Zhao et al., "Imaging of nonlinear and dynamic functional brain connectivity based on eeg recordings with the application on the diagnosis of Alzheimer's disease," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1571–1581, May 2020.
[3] H. Cui et al., "BrainGB: A benchmark for brain network analysis with graph neural networks," *IEEE Trans. Med. Imag.*, vol. 42, no. 2, pp. 493–506, Feb. 2023.
[4] M. Feng and J. Xu, "Detection of asd children through deep-learning application of fMRI," *Children*, vol. 10, no. 10, 2023, Art. no. 1654.

[5] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A survey on deep learning for neuroimaging-based brain disorder analysis," *Front. Neurosci.*, vol. 14, pp. 1–19, 2020.
[6] W. Yin, L. Li, and F.-X. Wu, "Deep learning for brain disorder diagnosis based on fMRI images," *Neurocomputing*, vol. 469, pp. 332–345, 2022.
[7] M. Saeidi et al., "Decoding task-based fmri data with graph neural networks, considering individual differences," *Brain Sci.*, vol. 12, no. 8, 2022, Art. no. 1094.
[8] N. Rasool and J. I. Bhat, "Unveiling the complexity of medical imaging through deep learning approaches," *Chaos Theory Appl.*, vol. 5, no. 4, pp. 267–280, 2023.
[9] M. P. McAvoy, L. Liu, R. Zhou, and B. A. Philip, "Reducing inter-individual differences in task fMRI preprocessing with OGRE (one-step general registration and extraction) preprocessing," *Neuroinformatics*, vol. 23, no. 47, pp. 1–12, Sep. 2025.
[10] H. Tang et al., "A comprehensive survey of complex brain network representation," *Meta- Radiol.*, vol. 1, no. 3, 2023, Art. no. 100046.
[11] K. Greff, R.K. Srivastava, J. Koutnık, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.
[12] Y. Liu et al., "A survey of visual transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7478–7498, Jun. 2024.
[13] Z. Wan, M. Li, S. Liu, J. Huang, H. Tan, and W. Duan, "EEFformer: A transformer–based brain activity classification method using EEG signal," *Front. Neurosci.*, vol. 17, 2023, Art. no. 1148855.
[14] P. Kim et al., "Swift: Swin 4 d fmri transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, vol. 36, pp. 42015–42037 .
[15] J. A. Nielsen et al., "Multisite functional connectivity mri classification of autism: Abide results," *Front. Hum. Neurosci.*, vol. 7, 2013, Art. no. 599.
[16] M. M. Rahman et al., "Interpreting models interpreting brain dynamics," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 12023.
[17] W. Wang et al., "Crossformer : A versatile vision transformer hinging on cross-scale attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3123–3136, May 2024.
[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30 .
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[20] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[21] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technologies*, vol. 2 (Short Papers), 2018, pp. 464–468.
[22] A.-. consortium, "The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience," *Front. Syst. Neurosci.*, vol. 6, 2012, Art. no. 62.
[23] C. R. Jack Jr et al., "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *J. Magn. Reson. Imaging: An Official J. Int. Soc. Magn. Reson. Med.*, vol. 27, no. 4, pp. 685–691, 2008.
[24] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, "The neuro bureau adhd-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275–286, 2017.
[25] T. Saramaki, "A class of linear-phase fir filters for decimation, interpolation, and narrow-band filtering," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-32, no. 5, pp. 1023–1036, Oct. 1984.
[26] J. Muschelli, M. B. Nebel, B. S. Caffo, A. D. Barber, J. J. Pekar, and S. H. Mostofsky, "Reduction of motion-related artifacts in resting state fmri using acompcor," *Neuroimage*, vol. 96, pp. 22–35, 2014.
[27] T. Sherif et al., "Cbrain: A web-based, distributed computing platform for collaborative neuroimaging research," *Front. Neuroinform.*, vol. 8, 2014, Art. no. 54.
[28] E. T. Rolls, C.-C. Huang, C.-P. Lin, J. Feng, and M. Joliot, "Automated anatomical labelling atlas 3," *Neuroimage*, vol. 206, 2020, Art. no. 116189.
[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[30] J. O. Caro et al., "BrainLM: A foundation model for brain activity recordings," in *Proc. Twelfth Int. Conf. Learn. Representations*, 2024, pp. 1–12.
[31] A. Thomas, C. Ré, and R. Poldrack, "Self-supervised learning of brain dynamics from broad neuroimaging data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 21255–21269.
[32] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3 d CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural mri," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.
[33] X. Kan, W. Dai, H. Cui, Z. Zhang, Y. Guo, and C. Yang, "Brain network transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 25586–25599.
[34] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler, "Self-supervised transformers for fmri representation," in *Proc. Int. Conf. Med. Imag. With Deep Learn.*,, 2022, pp. 895–913.
[35] J. Park, C. Kim, and W. Sung, "Convolution-based attention model with positional encoding for streaming speech recognition on embedded devices," in *Proc. 2021 IEEE Spoken Lang. Technol. Workshop (SLT)*, 2021, pp. 30–37.
[36] W. Park, W. Chang, D. Lee, and J. Kim, "Grpe: Relative positional encoding for graph transformer," in *Proc. ICLR Mach. Learn. Drug Discovery*, 2022, pp. 1–12.
[37] N. Kokhlikyan et al., "Captum: A unified and generic model interpretability library for pytorch," 2020, *arXiv:2009.07896*.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[39] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.
[40] T. Chen et al., "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," *ACM SIGARCH Comput. Archit. News*, vol. 42, no. 1, pp. 269–284, 2014.
[41] H. Zhang, G. Li, J. Li, Z. Zhang, Y. Zhu, and Z. Jin, "Fine-tuning pre-trained language models effectively by optimizing subnetworks adaptively," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 21442–21 454 .
[42] S. A. Salman, Z. Lian, M. Saleem, and Y. Zhang, "Functional connectivity based classification of adhd using different atlases," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, 2020, pp. 62–66.
[43] Y. Paloyelis, M. A. Mehta, J. Kuntsi, and P. Asherson, "Functional mri in adhd: A systematic literature review," *Expert Rev. Neurotherapeutics*, vol. 7, no. 10, pp. 1337–1356, 2007.
[44] D. Mousa, N. Zayed, and I. A. Yassine, "Alzheimer disease stages identification based on correlation transfer function system using resting-state functional magnetic resonance imaging," *Plos One*, vol. 17, no. 4, 2022, Art. no. e0264710.
[45] X.-a. Bi, Q. Jiang, Q. Sun, Q. Shu, and Y. Liu, "Analysis of Alzheimer's disease based on the random neural network cluster in fmri," *Front. Neuroinform.*, vol. 12, 2018, Art. no. 60.
[46] N. L. Faizo, "A narrative review of mri changes correlated to signs and symptoms of autism," *Medicine*, vol. 101, no. 34, 2022, Art. no. e30059.
[47] Y. ElNakieb et al., "Understanding the role of connectivity dynamics of resting-state functional mri in the diagnosis of autism spectrum disorder: A comprehensive study," *Bioengineering*, vol. 10, no. 1, 2023, Art. no. 56.