



## Efficient self-attention mechanism and structural distilling model for Alzheimer's disease diagnosis

Jiayi Zhu<sup>a</sup>, Ying Tan<sup>a,\*</sup>, Rude Lin<sup>a</sup>, Jiaqing Miao<sup>a,\*</sup>, Xuwei Fan<sup>a</sup>, Yafei Zhu<sup>a</sup>, Ping Liang<sup>a</sup>, Jinnan Gong<sup>b,c</sup>, Hui He<sup>c</sup>

<sup>a</sup> The Key Laboratory for Computer Systems of State Ethnic Affairs Commission, Southwest Minzu University, Chengdu, 610041, China

<sup>b</sup> School of Computer Science, Chengdu University of Information Technology, Chengdu, China

<sup>c</sup> The Key Laboratory for NeuroInformation of Ministry of Education, High-Field Magnetic Resonance Brain Imaging Key Laboratory of Sichuan Province, Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu, 610054, China

### ARTICLE INFO

#### Keywords:

Alzheimer's disease  
Magnetic resonance imaging  
Classification  
Self-attention mechanism  
Feature distilling  
Computational complexity

### ABSTRACT

Structural magnetic resonance imaging (sMRI) is commonly used for the identification of Alzheimer's disease because of its keen insight into atrophy-induced changes in brain structure. Current mainstream convolutional neural network-based deep learning methods ignore the long-term dependencies between voxels; thus, it is challenging to learn the global features of sMRI data. In this study, an advanced deep learning architecture called Brain Informer (BraInf) was developed based on an efficient self-attention mechanism. The proposed model integrates representation learning, feature distilling, and classifier modeling into a unified framework. First, the proposed model uses a multihead ProbSparse self-attention block for representation learning. This self-attention mechanism selects the first  $\lfloor \ln N \rfloor$  elements that can represent the overall features from the perspective of probability sparsity, which significantly reduces computational cost. Subsequently, a structural distilling block is proposed that applies the concept of patch merging to the distilling operation. The block reduces the size of the three-dimensional tensor and further lowers the memory cost while preserving the original data as much as possible. Thus, there was a significant improvement in the space complexity. Finally, the feature vector was projected into the classification target space for disease prediction. The effectiveness of the proposed model was validated using the Alzheimer's Disease Neuroimaging Initiative dataset. The model achieved 97.97% and 91.89% accuracy on Alzheimer's disease and mild cognitive impairment classification tasks, respectively. The experimental results also demonstrate that the proposed framework outperforms several state-of-the-art methods.

### 1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that worsens over time and is accompanied by irreversible neuronal damage with progressive impairment of cognitive functions [1,2]. Mild cognitive impairment (MCI) is the prodromal stage of AD, and patients in this state have a high probability of developing AD [3]. In recent years, with the increasing number of patients with AD and MCI, the need for computer-aided-diagnosis has dramatically increased [4–6]. The structural magnetic resonance imaging (sMRI) technique provides powerful data to support this need because of its ability to noninvasively capture structural changes in the brain caused by the atrophic process [7,8]. Fan et al. [9] used a support vector machine (SVM) to classify and predict different disease processes in AD using sMRI data. Lian et al. [10] used whole-brain sMRI data for AD diagnosis using joint learning and multiscale feature representation. Kumar et al. [11] used

transfer learning to classify AD and normal controls (NC) after entropy slicing to select the most informative sMRI slices. Khatri et al. [12] used multiple biomarkers obtained from sMRI processing for the classification of AD and MCI. Odusami et al. [13,14] applied multiple deep feature extractors to enhance the classification performance of multiple AD stages and proposed a fine-tuned deep learning network for further performance improvement. Razzak et al. [15] performed a feature fusion of different convolutional kernel sizes on sMRI data for the classification of AD, MCI, and NC. Ashraf et al. [16] used a transfer learning strategy for neurological disorder detection based on sMRI data. A large number of similar studies have shown that structural abnormalities in the brain are closely associated with brain disorders, and the feasibility of using sMRI data for AD classification has been demonstrated.

\* Corresponding authors.

E-mail addresses: [ty7499@swun.edu.cn](mailto:ty7499@swun.edu.cn) (Y. Tan), [jiaqing\\_miao@swun.edu.cn](mailto:jiaqing_miao@swun.edu.cn) (J. Miao).

<https://doi.org/10.1016/j.complbiomed.2022.105737>

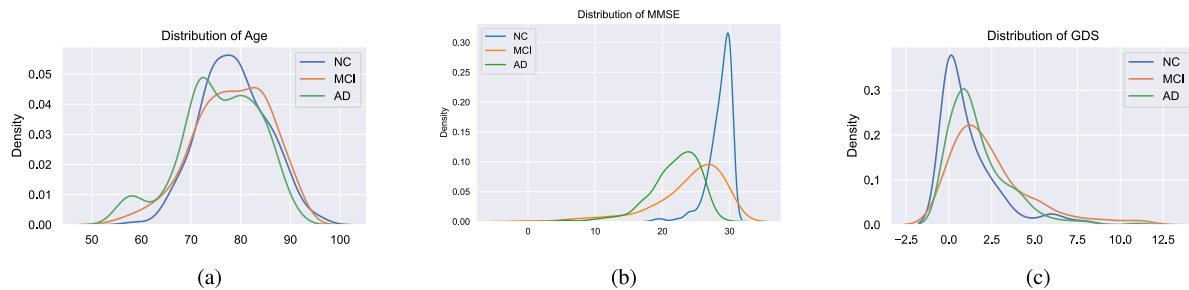
Received 2 April 2022; Received in revised form 23 May 2022; Accepted 11 June 2022

Available online 17 June 2022

0010-4825/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Table 1**  
Demographic and clinical information of the dataset.

Type	Sex (male/female)	Age (mean $\pm$ std)	MMSE (mean $\pm$ std)	GDS (mean $\pm$ std)
NC	198/126	79.35 $\pm$ 5.29	28.96 $\pm$ 1.31	0.99 $\pm$ 1.22
MCI	239/80	77.21 $\pm$ 6.99	26.13 $\pm$ 3.16	1.70 $\pm$ 1.76
AD	163/153	75.70 $\pm$ 7.88	21.88 $\pm$ 3.70	1.69 $\pm$ 1.76



**Fig. 1.** The density maps of each indicator in the dataset. (a) Age distribution; (b) MMSE score distribution; (c) GDS distribution. Comparison of the three shows that GDS and MMSE scores are more clearly distinguished in each subject type, and age is more similarly distributed.

Deep learning methods are gradually becoming mainstream in the current artificial intelligence field and are dominated by convolutional neural networks (CNNs) [17,18]. Lee et al. [19] introduced a framework for sMRI classification of AD using AlexNet [20]. Zhang et al. [21] proposed a lightweight network based on the ResNet architecture [22] and used sMRI data for AD/NC classification. Ahsan et al. [23] constructed a multiple two-dimensional CNN network to learn the local features of sMRI data for AD classification. Since convolution operations are performed using fix-sized kernels [24] (e.g.,  $3 \times 3$  size convolution kernels), this leads to the fact that the learned features can only focus on local regions of the brain, and distant features across brain regions and global features are limited [25]. Therefore, the improvement of CNNs applied to MRI is hindered.

While CNNs were developing, a model called Transformer caused a sensation in the field of natural language processing (NLP) [26]. The core operation of this model is the self-attention mechanism that emphasizes the dependencies between long sequences. Thus, it is superior to learning global representation. The excellent performance of this mechanism has led many researchers to attempt its migration to the computer vision (CV) domain. For example, Detection Transformer (DETR) [27] uses Transformer for object detection; Vision Transformer (ViT) [28] introduces the idea of segmenting an image into several patches for processing, with outstanding classification results after training on a large dataset; Swin-Transformer [29] uses a local self-attention mechanism on top of ViT to reduce the calculation complexity; Transformer-in-Transformer (TNT) [30] models both local and global features of an image to retain spatial information. However, because of the  $O(n^2)$  complexity of the calculations within this mechanism [31], a huge computational overhead would be incurred if it is directly applied to MRI data. Therefore, the application of the self-attention mechanism in MRI is somewhat limited.

To overcome these limitations, the Brain Informer (BraInf) model is proposed in this study for efficient feature encoding and data classification of MRI data. First, considering the three-dimensional tensor data, a feature extraction strategy is designed to convert the raw three-dimensional MRI data into a two-dimensional feature matrix. Second, to lower the quadratic complexity of the original self-attention mechanism, the multihead ProbSparse self-attention mechanism [32] is used in the proposed framework to improve the operation efficiency. The mechanism selects only the most informative first  $\lfloor \ln N \rfloor$  elements from the perspective of probability sparsity to represent overall features. Therefore, the computational complexity is reduced, and high performance is guaranteed simultaneously. Third, to ensure network depth while maintaining the memory cost at a low level, structural distilling is proposed which is a module that merges tensor patches to

gradually reduce the size of the feature map. Finally, the output from the framework was passed into the classifier for disease classification. We validated the proposed framework on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and demonstrated the superiority of the method in terms of algorithm performance and various medical metrics such as accuracy, specificity, sensitivity, and precision.

The major contributions of the study are as follows:

1. The proposed structural distilling applied the idea of patch merging into the distilling operation without maxpooling, which reduces the information loss in the downsampling process and improves performance. Thus, it lowers the space cost for calculation while preserving as many features as possible, thereby making the network deeper.
2. The proposed architecture significantly reduces the computational complexity compared with the original self-attention models, thus making larger-scale models trainable on large datasets.
3. This study uses original three-dimensional sMRI data, which is a data-driven method that does not rely on prior knowledge, and its accuracy can compete with state-of-the-art methods.

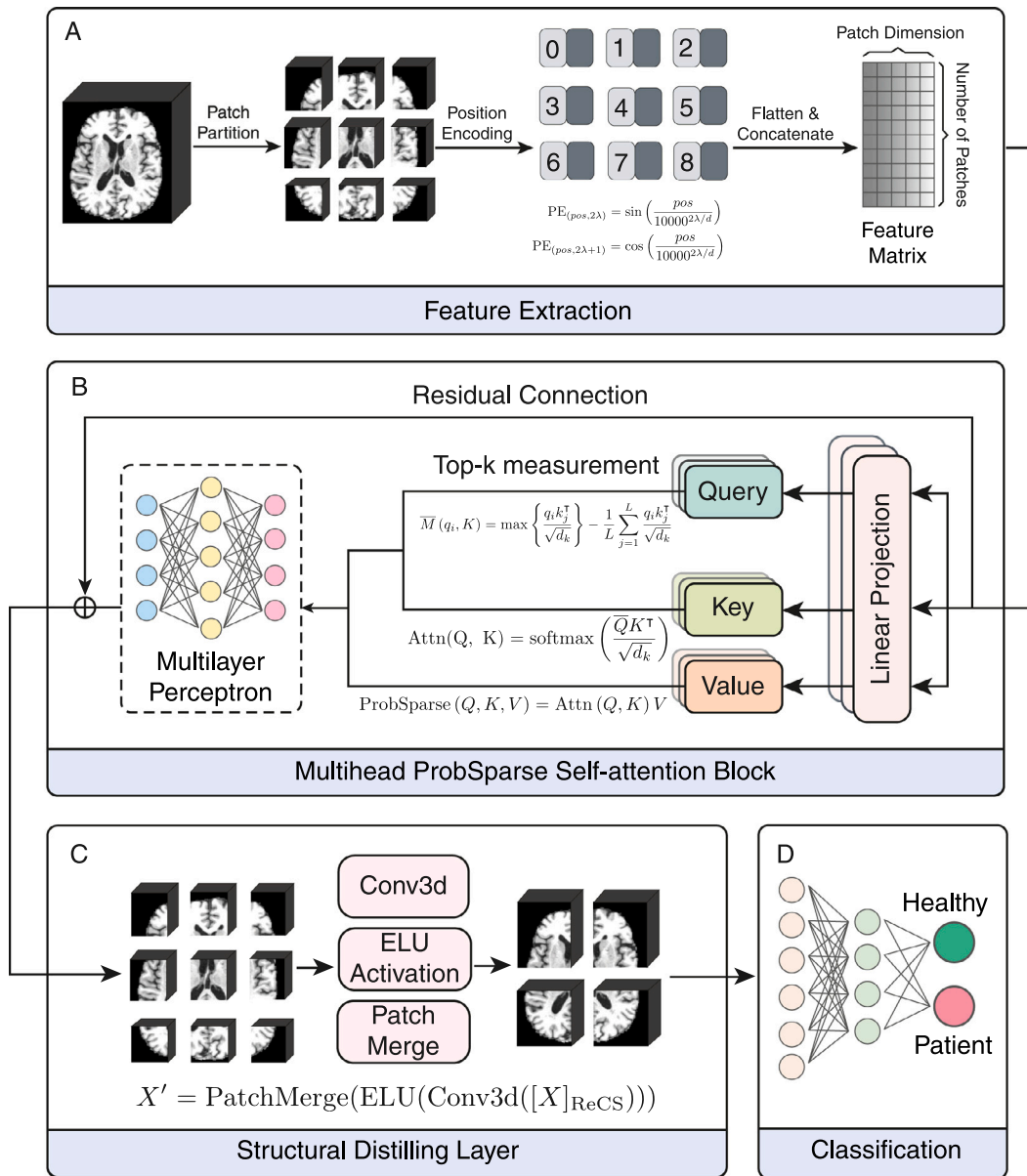
The remainder of this paper is organized as follows. Section 2 briefly introduces the datasets used in this study. Section 3 provides a detailed description of the proposed framework. Section 4 presents various comparative experiments on the model in detail, as well as a discussion of the experimental results. Finally, Section 5 concludes the study.

## 2. Materials

In this section, we introduce the MRI dataset and the preprocessing pipeline used in this study.

### 2.1. Studied dataset

The dataset used for this study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, available at <http://adni.loni.usc.edu/>) database. The ADNI provides scientists worldwide with a publicly available AD database to enable the early diagnosis of AD and the exploration of biological indicators of the disease [33]. In this study, we selected 324 NC, 319 MCI, and 316 AD samples. All MRI data were obtained using 3T scanning equipment, and these data were T1-weighted images with magnetization-prepared rapid-acquisition gradient-echo sequences. The sequence parameters were as follows: field of view (FOV) =  $208 \times 240 \times 256$  mm, resolution =



**Fig. 2.** Overall flowchart of the BraInf architecture. (A) Feature extraction phase. The original three-dimensional sMRI data is first patch partitioned with fixed patch size. The blocks are positional encoded. After adding the original block with positional encoded information, each block is flattened into a vector. All the vectors are concatenated into a matrix as the feature matrix. (B) Multihead ProbSparse self-attention block. The feature matrix is formed into query, key, and value matrices after multiple linear projections, and the ProbSparse self-attention is computed using the formula in (B). The multilayer perceptron is applied for further representation learning of the data, followed by a residual connection. (C) Structural distilling layer. The feature matrix is first reconstructed into three-dimensional blocks. Then Conv3d (·), ELU (·), and PatchMerge (·) are applied to obtain the down-sampled three-dimensional block. It reduced the feature scale and lowered the space complexity of the model. (B) and (C) can be stacked multiple times, to make the network deeper. (D) Classifier (Feedforward neural network). The subject is classified as healthy or patient after passing the feature vector into the classifier.

1 × 1 × 1 mm, repetition time (TR) = 2,300 ms, and inversion time (TI) = 90 ms.

Demographic and clinical characteristics, including sex, age, Mini-Mental State Examination (MMSE) scores, and Geriatric Depression Scale (GDS), are shown in Table 1. The density maps for each indicator in the dataset is shown in Fig. 1. The ages of the three subject types were distributed similarly. The GDS and MMSE scores of the NC group showed low variance, whereas the other two groups had relatively high variance.

2.2. Data preprocessing

First, all raw sMRI data were relocated to the midpoint of the anterior commissure (AC) - posterior commissure (PC) line. Then, we used the computational anatomy toolbox (CAT12, available at <http://www.neuro.uni-jena.de/cat/>) for SPM [34], which covers a variety of morphometry methods, such as voxel-based morphometry (VBM) [35] and surface-based morphometry (SBM) [36]. The following steps were performed in our preprocessing steps: (1) non-brain tissue removal, including the skull and neck, etc. (2) normalization to the EPI template, (3) modulation, and (4) spatial smoothing using a Gaussian filter of 8 mm full-width at half-maximum (FWHM).

3. Methods

The overall architecture proposed in this study is illustrated in Fig. 2. First, a patch partition operation is performed on the MRI data [28], which equally divides the original image into multiple blocks. Each block is then linearly transformed, and the original position encoding is added. Finally, these blocks were sequentially arranged

into a matrix and passed into the main model as the feature matrix of the sample. The feature matrix is first passed to the multihead ProbSparse self-attention block [32]. Compared to traditional deep learning methods, the self-attention mechanism can effectively extract the correlation between long-range sequences and capture more comprehensive features [28]. The output of the self-attention block was passed to the multilayer perceptron for further representation learning [37]. After the residual connection, it is followed by the structural distilling layer, which is a computational module that reduces the size of the feature map while retaining the key features of the original data, thereby reducing the computational cost. The structural distilling layer enables the model to handle large-scale input features. The self-attention mechanism and structural distilling layer can be superimposed multiple times [26], allowing a deeper model structure. Finally, the output of the model was passed into the classifier for disease prediction.

### 3.1. Related work

#### 3.1.1. Self-attention mechanism

Vaswani et al. [26] first proposed the self-attention mechanism. In the field of NLP, it is usually implemented with recurrent neural networks [38], also known as encoder–decoder structures [39]. Because this type of model uses sequence-by-sequence processing, it is difficult to extract the long-range dependencies. In addition, long sequences lead to “gradient vanishing” or “gradient explosion” during training, which makes it difficult for the network to converge. A later study attempted to incorporate the attention mechanism into the encoder–decoder structure [40]. This study used weighted summation to compute attention so that the neurons in the decoder focused on the key parts of the encoder content. Although the performance of the model was improved by the attention mechanism to some extent, the backbone network was still a recurrent neural network, and the fundamental defects of the encoder–decoder structure were still not solved. The emergence of the self-attention mechanism has solved the shortcomings of the encoder–decoder structure and has become a mainstream approach for NLP.

However, the biggest problem with the self-attention mechanism is its quadratic computational complexity, which makes it difficult to handle large amounts of input data. Some studies have attempted to solve this issue, but they still have limitations. Sparse Transformer [41] reduces the time complexity to  $O(n\sqrt{n})$  based on sparse factorizations of the attention matrix, but the efficiency improvement is limited. Reformer [42] used a locality-sensitive hashing technique to reduce complexity as well, but it shows a performance improvement only for extremely long input sizes. Transformer-XL [43] further proposed a segment-level recurrence mechanism that enables longer-term dependency but is not conducive to breaking the efficiency bottleneck. The problems of high computational complexity and performance bottleneck breakthroughs remain unresolved.

#### 3.1.2. Distilling operation

In addition to the computational complexity problem, another issue of deep learning models based on the self-attention mechanism is the memory bottleneck and feature redundancy. In self-attention models, multiple self-attention blocks are typically stacked to deepen the model structure [26]. However, stacking attention layers causes redundancy in the feature representation [32]. In addition, self-attention calculates each of the two input tokens, thus requiring  $O(n^2)$  space complexity. The memory usage of  $O(J \cdot n^2)$  is required to superimpose  $J$  times the self-attention blocks. In MRI processing, it can easily yield a large  $n$  value. Therefore, owing to the limitation of memory, a large  $J$  value cannot be set in practice, resulting in a shallow model and an inability to extract deeper features of the data. The distilling operation [32] solves the above issues: (1) Distilling refines the output features of self-attention. The refined feature map contains the most informative sequences, thus reducing the performance degradation caused by the

redundancy of internal calculations. (2) By reducing a certain amount of sequence information by distilling, the model can handle data with large  $n$  values well. This made it possible to build a deeper model. However, it does not consider the three-dimensional MRI structure. The operation will cause the destruction of the three-dimensional structure if it is applied directly to the MRI data. Therefore, the feature representations of the MRI data are less informative. Furthermore, the maxpooling operation inside the distilling may cause data loss and reduce the feature learning capability of the model to some extent.

### 3.2. Feature extraction

#### 3.2.1. Patch partition

In this study,  $X_p \in \mathbb{R}^{L \times W \times H}$  denotes the original three-dimensional sMRI image, and  $P \in \mathbb{R}^{P_L \times P_W \times P_H}$  denotes a patch tensor. First, we divide the original image into  $N$  patches, where  $N = \frac{L}{P_L} \cdot \frac{W}{P_W} \cdot \frac{H}{P_H}$ . Then, all patches are flattened and transformed into vectors  $x_p^i \in \mathbb{R}^{1 \times d}$ , where  $i$  represents the  $i$ th patch and  $d = P_L \cdot P_W \cdot P_H$  is the dimension of the patch tensor. Finally, we concatenate all patch vectors to construct a matrix  $X_p \in \mathbb{R}^{N \times d}$  for further steps.

#### 3.2.2. Positional encoding

One drawback of using the self-attention mechanism in processing sequential data is that all tokens are passed into the model simultaneously, resulting in an inability to retain the positional information of each token. Therefore, it is crucial to encode the position information for each token. For the sMRI data, each input token is an individual patch after the patch partition. Here, the position information is encoded using sine and cosine functions to label each patch and construct the position relationship between them [26], defined as:

$$PE_{(pos,2\lambda)} = \sin(pos/10000^{2\lambda/d}), \quad (1)$$

$$PE_{(pos,2\lambda+1)} = \cos(pos/10000^{2\lambda/d}), \quad (2)$$

where  $pos$  denotes the position of the current patch,  $\lambda$  denotes the  $\lambda$ -th dimension and  $d$  denotes the patch dimension. Each dimension of the position encoding corresponds to a sine curve with a waveform varying from  $2\pi$  to  $10000 \cdot 2\pi$ . The trigonometric function is used to describe the position information because it can convert the relative positions between the patches into absolute positions, that is, each  $PE_{pos+k}$  can be obtained by a linear transformation of  $PE_{pos}$ .

Finally, the inputs are fused with the position information to obtain the final feature matrix:

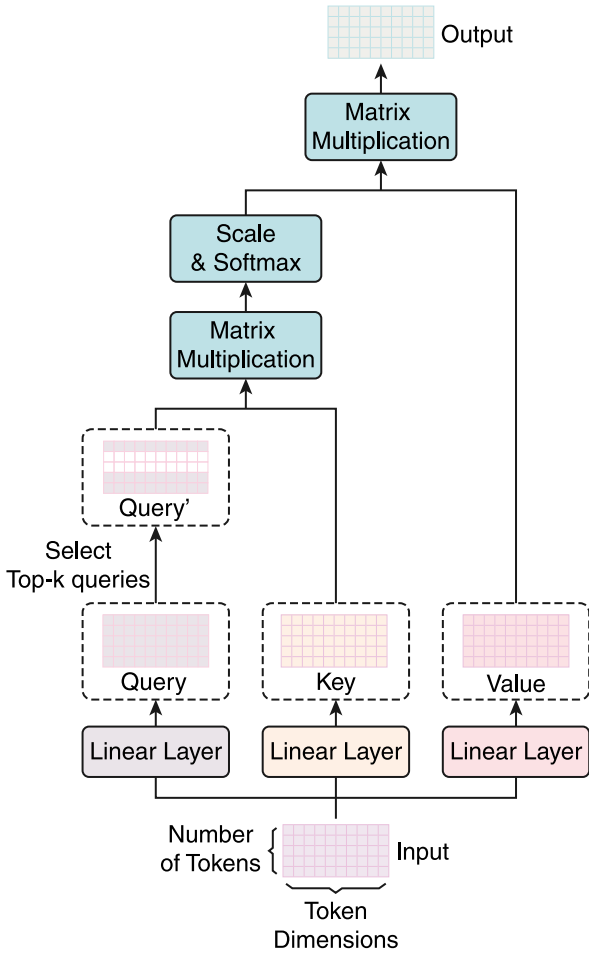
$$X = X_p + PE, PE \in \mathbb{R}^{\lambda \times d}, \quad (3)$$

where  $X$  represents the feature matrix to be input into the model after patch partitioning and positional encoding.

### 3.3. Multihead ProbSparse self-attention block

Zhou et al. [32] proposed the ProbSparse self-attention mechanism to address some of the shortcomings of the original self-attention mechanism. Most notably, it was designed to solve the high computational complexity from  $O(n^2)$  to  $O(n \log n)$ . The performance of this mechanism in long sequence prediction tasks is significantly improved compared with that of conventional self-attention. The flow of the ProbSparse self-attention mechanism is illustrated in Fig. 3. Suppose that  $L$  is the length of the sequence and  $d$  is the dimension of each sequence. The input matrix  $X \in \mathbb{R}^{L \times d}$  is first transformed into three different matrices by three different linear layers: query ( $Q$ ), key ( $K$ ), and value ( $V$ ).

$$\begin{aligned} Q &= XW_q, \\ K &= XW_k, \\ V &= XW_v, \end{aligned} \quad (4)$$



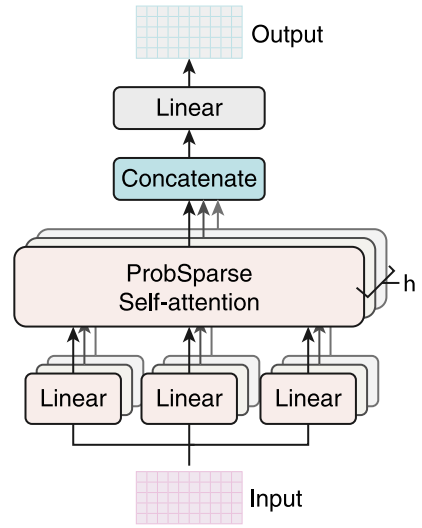
**Fig. 3.** ProbSparse self-attention mechanism. The input matrix is transformed from three trainable matrices into three matrices  $Q, K, V$ . The top- $k$  queries with the highest influence are selected for the  $Q$  matrix, and the other queries are replaced by zero vectors. The blank part of  $Q'$  ( $\bar{Q}$ ) represents the filtered queries, and the retained part remains unchanged. Finally, it is multiplied with the  $V$  matrix to obtain the output of ProbSparse self-attention.

where  $W_q, W_k$  and  $W_v$  are trainable weight matrices, and  $W_q \in \mathbb{R}^{d \times d_k}$ ,  $W_k \in \mathbb{R}^{d \times d_k}$  and  $W_v \in \mathbb{R}^{d \times d_v}$ . We then calculate the similarity between each query in  $Q$  and each key in  $K$  using the dot-product. The ProbSparse self-attention utilizes probability sparsity to select the most informative queries for the computation. The probability distribution of the attention map forms a “long-tail” distribution, that is, only a few elements in the attention map play a decisive role. The “active” queries tend to contain these key elements, while other “lazy” queries do not. We eliminate these “lazy” queries (mask as 0) and retain the Top- $k$  informative queries. The new query matrix  $\bar{Q}$  was used for the following self-attention calculation:

For ease of presentation, we use  $q_i, k_i$  to represent the  $i$ th row, that is, the  $i$ th query and the  $i$ th key, respectively, in the  $Q, K$  matrices. According to [44], the probability form of the attention  $q_i$  to  $k_j$  is defined as:

$$\text{Attn}(q_i, k_j) = p(k_j|q_i) = \frac{\exp\left(\frac{q_i k_j^T}{\sqrt{d_k}}\right)}{\sum_l \exp\left(\frac{q_i k_l^T}{\sqrt{d_k}}\right)}. \quad (5)$$

Based on this probability form, it is possible to compare  $p(k_j|q_i)$  with the uniform distribution,  $q(k_j|q_i) = \frac{1}{L}$ . If  $p(k_j|q_i)$  approximates a uniform distribution, it means that  $p_i$  is likely to be a “lazy” query. We used the Kullback–Leibler divergence to measure the similarity of two



**Fig. 4.** Diagram of the multi-head self-attention mechanism. The input matrix is passed into multiple self-attention mechanisms to learn the representations of the data in parallel. The components are concatenated after computation and passed into a linear layer as the final output. The  $h$  in the figure indicates the number of heads.

probability distributions, and the definition of the importance of the  $i$ th query is defined as:

$$M(q_i, K) = \ln \sum_{j=1}^L \exp\left(\frac{q_i k_j^T}{\sqrt{d_k}}\right) - \frac{1}{L} \sum_{j=1}^L \frac{q_i k_j^T}{\sqrt{d_k}}. \quad (6)$$

To simplify the calculation, the above equation was modified in [32] as:

$$\bar{M}(q_i, K) = \max\left\{\frac{q_i k_j^T}{\sqrt{d_k}}\right\} - \frac{1}{L} \sum_{j=1}^L \frac{q_i k_j^T}{\sqrt{d_k}}. \quad (7)$$

According to this definition, the Top- $k$  queries with the highest values are retained, and the rest are filled with zeros to obtain the  $\bar{Q}$  matrix. Subsequently, a regular self-attention calculation was performed.

$$\text{Attn}(Q, K) = \text{softmax}\left(\frac{\bar{Q}K^T}{\sqrt{d_k}}\right), \quad (8)$$

$$\text{ProbSparse}(Q, K, V) = \text{Attn}(Q, K)V. \quad (9)$$

This process is described in Algorithm 1. The “multihead” approach of self-attention further enhances the ability to learn representations of the data, as shown in Fig. 4. Multihead self-attention is essentially a multi-directional learning of the input matrix with multiple sets of different self-attentions. These multiple sets of self-attention tasks do not interfere with each other. These self-attentions are computed in parallel, allowing the model to improve its performance with low computational complexity. After computing each attention head, these outputs are concatenated and passed into the final linear layer as the output of the multi-head self-attention mechanism. The above process can be described as follows:

$$MPSA(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W_O, \quad (10)$$

where  $\text{head}_i = \text{ProbSparse}(XW_i^Q, XW_i^K, XW_i^V)$  and  $W^Q, W^K, W^V, W_O$  are trainable matrices. Multihead self-attention prevents partial information loss due to single-head mapping and is therefore more likely to yield better training results.

### 3.4. Multilayer perceptron

Self-attention is immediately followed by a multilayer perceptron (MLP). The mere superposition of self-attention causes the model to

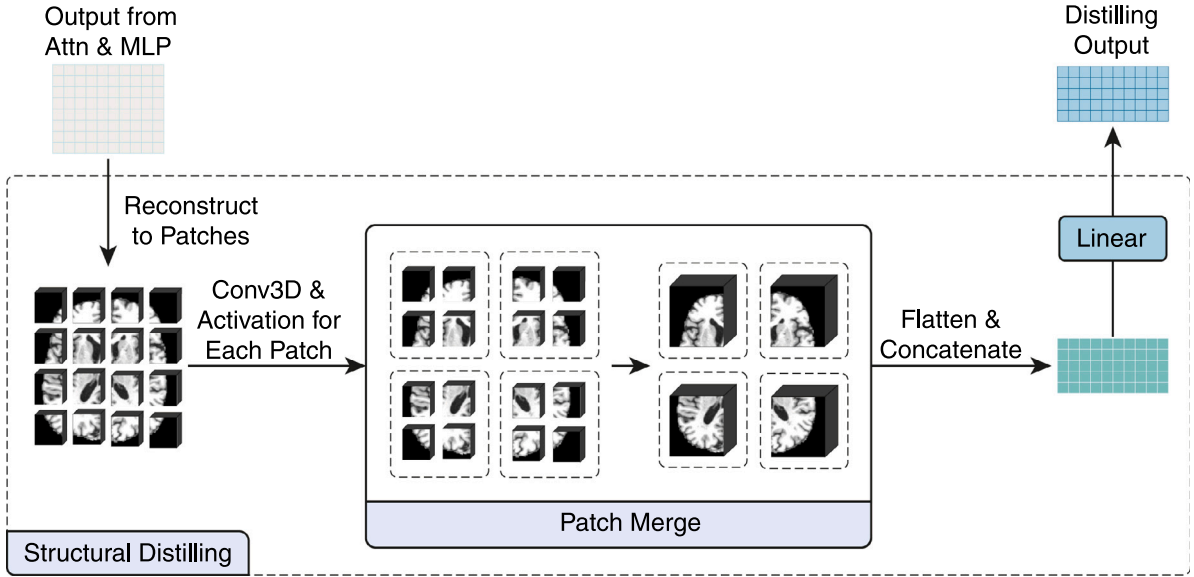


Fig. 5. Structural distilling operation. The output of the self-attention and MLP will be used as the input of distilling layer. The input matrix are first reconstructed to MRI patches, and then Conv3D and ELU activation are performed for each patch separately. In downsampling, more feature information is preserved by using patch merging for each four adjacent patches. The MRI patches after patch merging are then flattened and concatenated into a matrix, followed by a linear mapping layer as the output of distilling.

#### Algorithm 1 ProbSparse Self-attention Mechanism

**Require:** Tensor  $Q, K, V \in \mathbb{R}^{L \times d}$

- 1: Set hyperparameters  $k = \ln L, u = L \ln L$
- 2: Randomly select  $u$  vectors from  $K$  as  $\bar{K}$
- 3: Calculate  $\bar{S} = Q\bar{K}$
- 4: Calculate  $\max(\bar{S}_i) - \text{average}(\bar{S}_i)$  for every row in  $\bar{S}$
- 5: Sort in descending order and select Top- $k$  as  $\bar{Q}$
- 6: Calculate  $\text{Attn} = \text{softmax}(\bar{Q}K^T / \sqrt{d})V$
- 7: Replace Attn according to original indices

**Output:** Attention Map *ProbSparseAttn*

converge to a rank-1 matrix, resulting in identical tokens. The increase in non-linearity due to the multilayer perceptron allows the rank collapse phenomenon to be resolved [37]. The multilayer perceptron consists of two layers: linear transformation and ReLU nonlinear activation.

$$X = \text{ReLU}(W_2(\text{ReLU}(W_1 X))), \quad (11)$$

where  $W_1$  and  $W_2$  are trainable linear transformation matrices, and  $\text{ReLU}(\cdot)$  represents the ReLU non-linearity activation function. Thus, the ProbSparse self-attention mechanism and multilayer perceptron together form the multihead ProbSparse self-attention block.

#### 3.5. Residual connection

One of the main problems with deep neural networks is gradient vanishing. As the model layers deepen, the gradient during training is likely to converge to zero, making it difficult to train the model effectively. This can cause a performance bottleneck in the model. He et al. [22] proposed a novel residual connection architecture that breaks the performance bottleneck and makes it more effective for training deeper models. In the proposed Brainf architecture, the mathematical expression for the residual connection is defined as:

$$X' = \text{MLP}(\text{Attn}(X)) + X, \quad (12)$$

where  $\text{Attn}(\cdot)$  is the self-attention block, and  $\text{MLP}(\cdot)$  is the multilayer perceptron. Residual connection is applied to alleviate the gradient vanishing problem and thus enhance the representation learning capability of the model.

#### 3.6. Structural distilling operation

The conventional distilling process [32] is defined as follows:

$$X_{j+1} = \text{MaxPool}(\text{ELU}(\text{Conv1d}(X_j))), \quad (13)$$

where  $X_j$  is the output of the MLP.  $\text{Conv1d}(\cdot)$  performs a one-dimensional convolution along the sequence dimension and then activates it using the  $\text{ELU}(\cdot)$  function [45].  $\text{MaxPool}(\cdot)$  achieves down-sampling of the sequence and preserves the key sequence information. The length of the sequence is reduced from  $L$  to  $L/2$  after distilling.

Because the original distilling destroys the three-dimensional structure of the MRI data and the information loss caused by the max-pooling operation, a structural distilling operation is proposed here. Fig. 5 shows the detailed process of the proposed structural distilling operation. The feature matrix was passed to the distilling block after MLP. Because each row in the sMRI feature matrix represents a three-dimensional MRI patch, we first reconstructed the feature matrix  $X_j$  into patches. Next, a three-dimensional convolution operation was applied to each MRI patch for further representation learning. The kernel size was set to (3,3,3) with a padding of 1, followed by ELU activation. Because  $\text{MaxPool}(\cdot)$  in the conventional approach causes data loss and does not maximize feature representations, the patch merge concept [29] is applied here. It makes good use of all the spatial features of the data without causing data loss and achieves feature sequence down-sampling simultaneously. Each of the four adjacent patches was merged. The new patches are then flattened and concatenated into a matrix. Finally, a linear layer is applied for the dimensional transformation as the output of the structural distilling. This process can be described by the following equation:

$$X' = \text{PatchMerge}(\text{ELU}(\text{Conv3d}([X_j]_{\text{ReCS}}))), \quad (14)$$

$$X_{\text{out}} = \text{Flatten}(X') \cdot W_{\text{out}}, \quad (15)$$

where  $X_j \in \mathbb{R}^{L \times d}$ ,  $\text{Flatten}(X') \in \mathbb{R}^{\frac{L}{4} \times 4d}$ ,  $W_{\text{out}} \in \mathbb{R}^{4d \times d}$  and the final output  $X_{\text{out}} \in \mathbb{R}^{\frac{L}{4} \times d}$ .  $[\cdot]_{\text{ReCS}}$  denotes the reconstruction operation that converts the two-dimensional feature matrix into three-dimensional patches. The sequences of the feature matrix after distilling were reduced by a quarter while retaining the key sequence information. It enables the model to increase the depth of the network as much as possible with limited computational space to ensure the representation capability of the model.

**Table 2**  
Model parameter details.

Model block	Calculation	Parameters	N (Number)
Input	–	3D Tensor (61 × 73 × 61)	1
Feature extraction	Patch partition	Patch size: (5 × 3 × 5)	1
	Positional encoding	$PE_{(pos,2\lambda)} = \sin(pos/10000^{2\lambda/d})$	1
		$PE_{(pos,2\lambda+1)} = \cos(pos/10000^{2\lambda/d})$	
	Flatten & Concatenate	–	–
Feature matrix size: (3456 × 75)	–	–	
BraInf main architecture	Self-attention block	Multihead ProbSparse self-attention: $n_{heads} = 8$ Multilayer perceptron: $n_{neurons} = 4 \times 75$	3
	Structural distilling	Reconstruct Conv3D (kernel size: (3 × 3 × 3)) ELU activation Patch merge & flatten	
		Flatten	
Classification	Feed-Forward network	Fully connected layer ( $n_{neurons} = 1024$ )	1
		ReLU activation	
		Fully connected layer ( $n_{neurons} = 256$ )	
		ReLU activation Fully Connected Layer ( $n_{neurons} = n_{classes}$ )	
Output	–	Predicted label	–

### 3.7. Training details

The details of each computational block parameter are presented in Table 2. Our model was trained on the following platforms: AMD EPYC 7302 16-core Processor, GeForce RTX 3090 with 23G memory, 251G RAM.

## 4. Results and discussion

In this study, two binary classification tasks were set to evaluate model performance: (1) NC vs. AD and (2) NC vs. MCI. To make the results more robust, all models were trained using 10-Fold cross-validation. All samples in the dataset were divided into ten subsets, nine of which were selected each time as the training set and the remaining one as the test set. After training ten times, the mean and variance of each metric were calculated for the final evaluation.

### 4.1. Overall performance

#### 4.1.1. Performance comparison between different machine learning methods

First, we compared various machine-learning models with the proposed BraInf architecture. The mainstream image classification models thus far can be divided into the following three categories: (1) Basic machine learning models. These machine-learning-based classification models have better classification results on relatively simple datasets and require a shorter time to train; (2) Convolution-based deep learning models. CNNs are among the most popular methods for processing image data. The classification accuracies of the most advanced CNNs so far have exceeded human accuracy. (3) Self-attention classification models. Among these three classification approaches, we selected a few models with the best results. The basic machine learning models include (1) Gaussian naive Bayes (GNB) [46], (2) logistic regression (LR) [47], (3) decision tree (DT) [48], and (4) adaptive boosting (AdaBoost) [49]. The CNN-based models included (5) VGG16 [50], (6) VGG19 [50], and (7) ResNet-18 [22]. Self-attention-based models include (8) Vision Transformer (ViT) [28], (9) Reformer [42], and (10) the proposed BraInf architecture. The feature extraction of different models varied based on the model architecture. For basic machine-learning-based approaches, we flattened the MRI data into a vector as the input. For the CNN-based models, the original three-dimensional

MRI data were used, and all the operations within the model were replaced with the corresponding three-dimensional forms. The input of the self-attention-based models is described in Section 3.3.

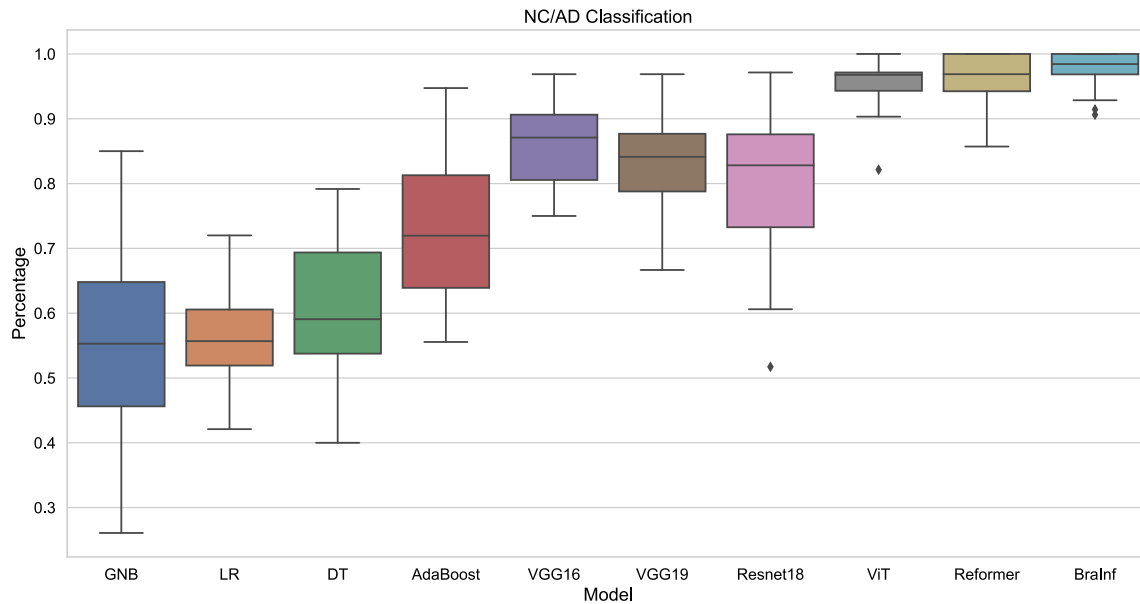
All models were validated on both AD/NC and MCI/NC classification tasks, and the classification performance was evaluated using four metrics: accuracy (ACC), sensitivity (SEN), specificity (SPE), and precision (PRE). Performance comparisons of each model in the two classification tasks are given in Table 3, Fig. 6, Table 4 and Fig. 7, respectively. The receiver operating characteristic (ROC) curves for the CNN-based and self-attention-based deep learning models are shown in Fig. 8.

In general, self-attention-based models have the best overall classification performance, followed by CNN-based deep learning models. The basic machine-learning models performed the worst. Due to the high dimensionality of MRI data, it is difficult for basic machine learning models to learn the representations of this particularly complex data, so the overall performance is not very impressive, with an average classification accuracy, sensitivity and specificity of 60.91%, 55.58%, and 66.50% in NC and AD classification tasks, and 58.36%, 58.47%, and 58.43% in NC and MCI, respectively. CNNs further improve the representation learning ability and exhibit excellent performance especially on a grid-like topology. Because MRI data are three-dimensional tensors, CNNs can also perform well on this data modality. The average accuracy, sensitivity, and specificity of the three CNNs in our experiments were 83.02%, 80.74%, and 85.39% for NC and AD, and 78.38%, 72.74%, and 83.74% for NC and MCI, respectively. Finally, self-attention-based models further improve performance because of their better ability to learn the global features of the data. The average accuracy, sensitivity, and specificity were 96.62%, 95.62%, and 97.56% for NC and AD, and 88.12%, 88.23%, and 88.07% for NC and MCI, respectively. It can be demonstrated that these self-attention models are superior to the machine learning and CNN-based approaches. For both classification tasks, the comprehensive performance of the BraInf model in this study was the best, and our model had the highest area under the curve (AUC), as shown in Fig. 8. This indicates that the model can effectively learn the representations of the MRI data.

Furthermore, a statistical analysis was performed to discuss the effectiveness of the proposed model. To compare the classification performance of several different models, a non-parametric Friedman test [51] was conducted to verify the significant differences in the various model performances. Subsequently, the post-hoc Wilcoxon method [52] was

**Table 3**  
NC/AD classification results of different models.

Classification framework	Model	ACC (mean $\pm$ std, %)	SEN (mean $\pm$ std, %)	SPE (mean $\pm$ std, %)	PRE (mean $\pm$ std, %)
Machine learning classification methods	GNB	54.78 $\pm$ 6.99	41.30 $\pm$ 8.43	68.84 $\pm$ 12.39	57.65 $\pm$ 12.77
	LR	56.18 $\pm$ 5.00	55.61 $\pm$ 8.26	55.82 $\pm$ 6.15	55.41 $\pm$ 7.20
	DT	60.86 $\pm$ 7.24	58.10 $\pm$ 7.98	63.99 $\pm$ 11.95	62.34 $\pm$ 8.52
	AdaBoost	71.82 $\pm$ 6.79	67.30 $\pm$ 12.19	77.33 $\pm$ 8.56	74.40 $\pm$ 10.48
CNN-based deep learning methods	VGG16	86.25 $\pm$ 4.98	88.35 $\pm$ 4.12	84.20 $\pm$ 7.47	84.83 $\pm$ 6.64
	VGG19	83.91 $\pm$ 5.46	84.12 $\pm$ 7.41	83.96 $\pm$ 6.46	83.82 $\pm$ 6.32
	ResNet-18	78.91 $\pm$ 4.49	69.74 $\pm$ 8.33	88.00 $\pm$ 7.77	85.83 $\pm$ 6.97
Attention-based deep learning methods	ViT	95.47 $\pm$ 2.56	93.25 $\pm$ 4.76	97.54 $\pm$ 1.84	97.33 $\pm$ 2.03
	Reformer	96.41 $\pm$ 2.10	95.86 $\pm$ 4.45	96.98 $\pm$ 2.33	96.79 $\pm$ 2.45
	<b>BraInf</b>	<b>97.97 <math>\pm</math> 1.41</b>	<b>97.74 <math>\pm</math> 2.19</b>	<b>98.17 <math>\pm</math> 2.85</b>	<b>98.16 <math>\pm</math> 2.68</b>



**Fig. 6.** Box plots of all metrics for each model in NC/AD classification. Basic machine learning approaches (GNB, LR, DT and AdaBoost) performed worse than the other two types of models in general. AdaBoost performed the best among them. The performance of CNN-based models (VGG16, VGG19 and Resnet18) is improved over traditional machine learning models. The self-attention-based models (ViT, Reformer and BraInf) performed best in general. In this classification task, our BraInf architecture performs best.

used to make significant comparisons of the performance between the models. The statistical results are visualized using a critical difference diagram [51], as shown in Fig. 9. It can be observed that the BraInf model has the highest average ranking in both the AD/NC and MCI/NC classification tasks. In addition, the classification performance of the BraInf model is significantly different from that of all the other models, which proves the superiority of the proposed architecture.

#### 4.1.2. Performance comparison with state-of-the-art methods

To validate the superiority of the proposed architecture, it was compared with recent advanced methods for the classification and diagnosis of Alzheimer's disease. The results are presented in Table 5. A direct comparison of the accuracy of the models is not the most reasonable method because of some differences between the amount of data used, modality of the data, and feature extraction methods, but it does allow for a rough comparison to some extent. In both classification tasks, the proposed architecture showed results comparable to those of the current state-of-the-art architecture. This indicates that the self-attention mechanism can substantially outperform convolution-based deep learning methods owing to its superior ability to capture long-range dependencies.

#### 4.2. Ablation study of the proposed framework

The two key blocks of the proposed BraInf architecture are multi-head ProbSparse self-attention and structural distilling. In this section, we describe the following ablation studies to demonstrate the importance of the blocks proposed in this study.

##### 4.2.1. Improvement of self-attention computational complexity

As mentioned in Section 3.3, the ProbSparse self-attention mechanism reduces the time complexity from  $O(n^2)$  to  $O(n \log n)$  compared with the original method. This results in significant performance improvement when dealing with large-scale inputs. Both the time and memory usage were evaluated for these two self-attention mechanisms in comparison.

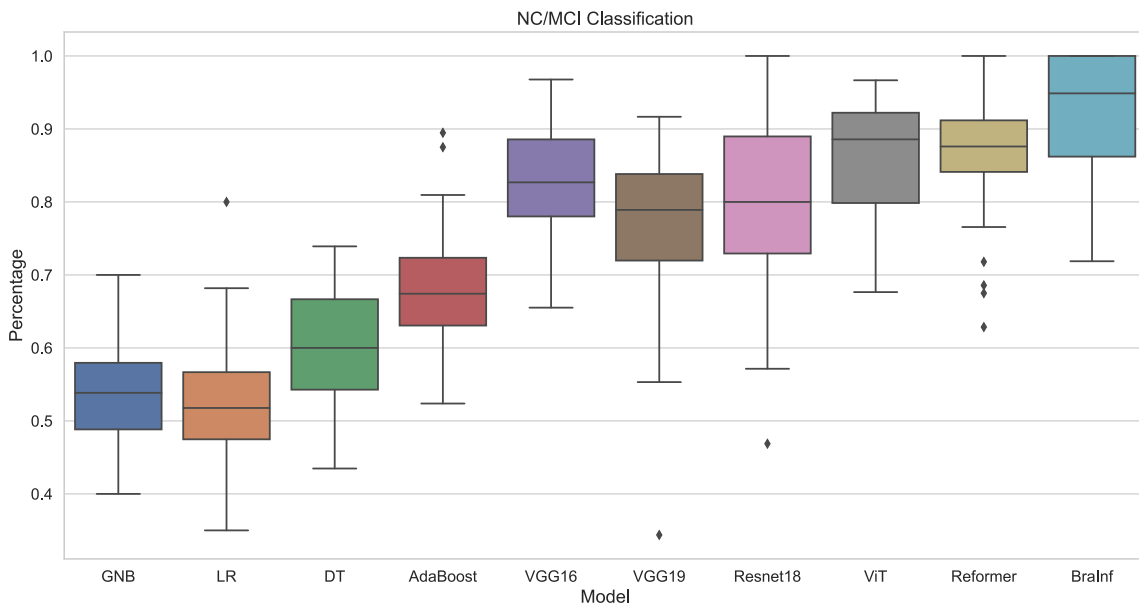
For sMRI data, the length of the sequence can be changed by different patch sizes; the smaller the patch size, the more sequences are generated, and vice-versa. A change in patch size also leads to a change in the sequence dimension. Various patch sizes also lead to changes in sequence dimensions. The effect of different patch sizes on the computation time was validated in our experiments. To compare only the effects of the attention mechanisms, the distilling layer was removed from the model, and the rest (multilayer perceptron, structure of feedforward neural network, etc.) was kept the same for both types of models.

The time and memory usage of the two self-attention mechanisms for different sequence lengths are presented in Table 6 and Fig. 10,

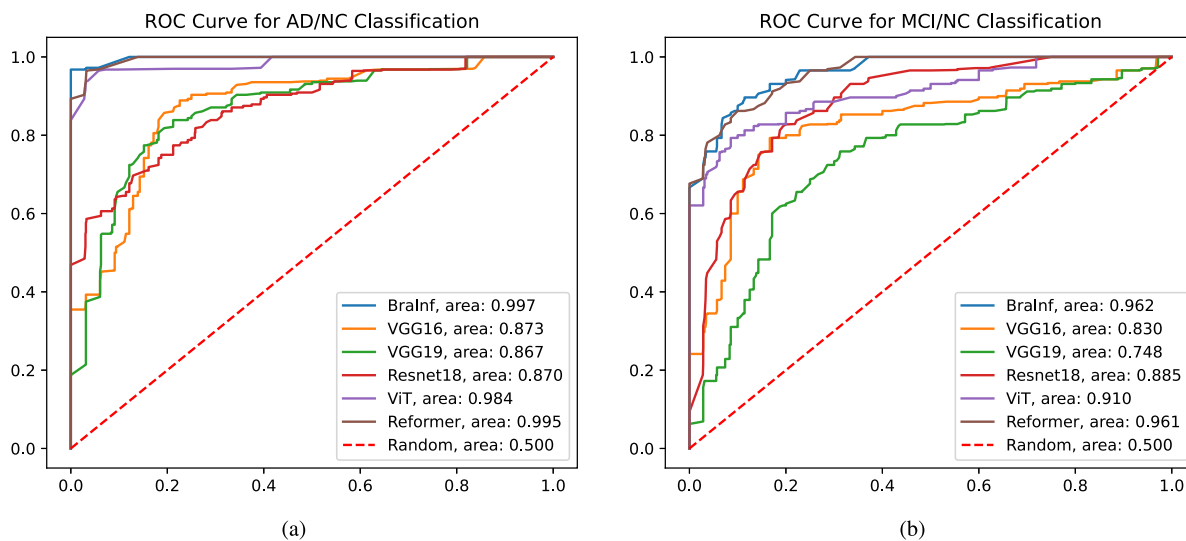


**Table 4**  
NC/MCI classification results of different models.

Classification framework	Model	ACC (mean ± std, %)	SEN (mean ± std, %)	SPE (mean ± std, %)	PRE (mean ± std, %)
Machine learning classification methods	GNB	53.28 ± 3.74	57.55 ± 6.55	49.40 ± 6.12	52.98 ± 6.74
	LR	52.33 ± 5.75	52.59 ± 12.33	51.69 ± 4.23	51.28 ± 9.77
	DT	60.07 ± 6.27	60.26 ± 8.88	59.94 ± 8.70	59.84 ± 8.36
	AdaBoost	67.76 ± 3.52	63.46 ± 7.90	72.68 ± 7.46	69.47 ± 9.89
CNN-based deep learning methods	VGG16	81.66 ± 5.92	77.04 ± 8.34	86.02 ± 6.40	84.06 ± 7.63
	VGG19	76.51 ± 7.93	76.96 ± 9.17	75.58 ± 16.44	76.97 ± 9.97
	ResNet-18	76.96 ± 6.08	64.22 ± 10.11	89.63 ± 6.27	85.95 ± 8.55
Attention-based deep learning methods	ViT	85.38 ± 6.24	82.56 ± 7.31	88.27 ± 7.71	87.15 ± 8.54
	Reformer	87.10 ± 5.68	<b>91.47 ± 4.94</b>	82.93 ± 9.96	84.48 ± 8.50
	<b>BraInf</b>	<b>91.89 ± 7.22</b>	90.66 ± 7.78	<b>93.01 ± 8.43</b>	<b>92.69 ± 8.21</b>



**Fig. 7.** Box plots of all metrics for each model in NC/MCI classification. Similar to NC/AD classification, conventional machine learning models showed low accuracy compared to other approaches. Self-attention-based architectures still performed best.



**Fig. 8.** ROC curve of CNN-based and self-attention-based models on two classification tasks. We use area under the curve (AUC) to measure the overall performance of the model. Larger area represents better performance. (a) ROC curve of AD/NC classification task. (b) ROC curve of MCI/NC classification task.

respectively. It can be observed that as the sequence length  $L$  increases, the growth in training time and memory usage of dot-product self-attention is much larger than that of ProbSparse self-attention. The

growth of dot-product shows a quadratic increase, while the growth of ProbSparse self-attention is flatter, thus proving that the ProbSparse mechanism can be well applied to data with a larger sequence size.

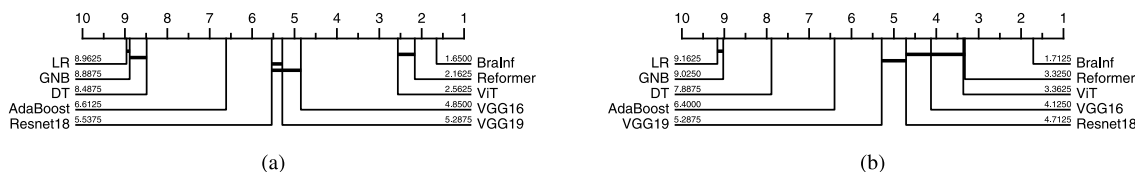


Fig. 9. Critical difference diagram of model performance on different classification tasks. Models with no significant difference in classification performance are connected by a short horizontal line. The number next to the model represents the average ranking of the corresponding model. Small values represent high ranking, which also represents better classification performance. (a) Diagram of AD/NC classification task. (b) Diagram of MCI/NC classification task.

Table 5 Performance comparison with state-of-the-art methods.

Methods	Subjects (NC/MCI/AD)	Modality	NC/AD classification			NC/MCI classification		
			ACC (%)	SEN (%)	SPE (%)	ACC (%)	SEN (%)	SPE (%)
Lian et al. (2020) [10]	429/-/358	MRI	90.3	82.4	96.5	-	-	-
Feng et al. (2020) [53]	200/280/200	MRI	94.2	96.6	92.4	84.6	89.7	77.5
Liu et al. (2020) [54]	119/233/97	MRI	88.9	86.6	90.8	76.2	79.5	69.8
Lin et al. (2021) [55]	707/-/649	MRI, PET	89.3	82.7	96.5	-	-	-
Ning et al. (2021) [56]	206/385/229	MRI, PET	96.9	95.7	98.0	82.1	87.1	72.3
Divya et al. (2021) [57]	347/558/171	MRI	96.8	92.8	98.8	89.4	95.2	80.1
Kang et al. (2021) [58]	229/382/187	MRI	90.4	93.9	83.8	72.4	74.7	84.4
Abdelaziz et al. (2021) [59]	226/226/186	MRI, PET, SNPs	98.2	97.8	98.8	93.1	92.7	93.6
Odusami et al. (2021) [14]	25/13/25	Functional MRI	80.8	91.8	83.9	92.2	90.2	94.2
Shanmugam et al. (2021) [60]	162/228/277	MRI	94.1	90.6	95.2	96.8	83.3	99.0
Goenka et al. (2022) [61]	475/224/70	MRI	98.4	94.0	-	97.7	96.0	-
Odusami et al. (2022) [13]	25/13/25	MRI	98.2 ACC, 98.1 SEN, 98.1 SPE (NC/AD/MCI)					
Li et al. (2022) [62]	330/299/299	MRI	93.2	95.0	89.8	80.4	83.2	78.6
Proposed method BraInf	324/316/319	MRI	98.0	97.7	98.2	91.9	90.1	93.0

Table 6 Time usage and memory usage of two self-attention mechanisms.

Patch size ( $L, d$ )	Dot-product attention model		ProbSparse attention model	
	Time (s) (mean $\pm$ std)	Memory (MiB) (mean $\pm$ std)	Time (s) (mean $\pm$ std)	Memory (MiB) (mean $\pm$ std)
[5 5 5] ( $L = 2016, d = 125$ )	0.398 $\pm$ 0.004	85.173 $\pm$ 1.404	0.492 $\pm$ 0.018	29.076 $\pm$ 1.065
[5 3 5] ( $L = 3456, d = 75$ )	0.655 $\pm$ 0.044	158.212 $\pm$ 0.846	0.551 $\pm$ 0.032	31.619 $\pm$ 0.484
[5 3 3] ( $L = 5760, d = 45$ )	1.160 $\pm$ 0.077	399.256 $\pm$ 0.455	0.498 $\pm$ 0.038	38.148 $\pm$ 1.024
[3 3 3] ( $L = 9600, d = 27$ )	2.349 $\pm$ 0.079	1073.105 $\pm$ 1.004	0.651 $\pm$ 0.046	45.651 $\pm$ 0.197
[3 2 3] ( $L = 14400, d = 18$ )	3.785 $\pm$ 0.136	2389.648 $\pm$ 2.276	0.684 $\pm$ 0.019	51.699 $\pm$ 1.374
[3 2 2] ( $L = 21600, d = 12$ )	6.917 $\pm$ 0.290	5355.717 $\pm$ 0.767	0.773 $\pm$ 0.062	66.172 $\pm$ 1.682
[2 2 2] ( $L = 32400, d = 8$ )	14.125 $\pm$ 0.219	12026.990 $\pm$ 3.598	0.881 $\pm$ 0.026	91.746 $\pm$ 4.197

Although the dimension  $d$  of the sequence can also affect the computational complexity to some extent, it is experimentally demonstrated that the computational complexity of self-attention mainly depends on the length of the sequence  $L$ .

#### 4.2.2. Contribution of structural distilling layer

In terms of spatial complexity, we analyzed in Section 3.6, that the superposition of multiple self-attention mechanisms can lead to a significant increase in memory usage, which limits the application of self-attention mechanisms in MRI research to some extent. The structural distilling operation ensures the accuracy of the model and

significantly reduces the memory usage, allowing self-attention to handle large-scale data. The performance improvement of the distilling operation was validated by pruning structural distilling in the model. Here, we evaluated the performance mainly in terms of memory usage. We conducted experiments with a batch size of 16 and patch size of  $5 \times 3 \times 5$ . The experimental results are presented in Tables 7 and 8, respectively. Table 7 shows the pruning of the distilling for the BraInf model. Table 8 shows the pruning of the distilling for the dot-product self-attention model. As shown in Table 7, the distilling operation substantially reduces the memory usage of the subsequent self-attention operations, whereas the overall memory usage without

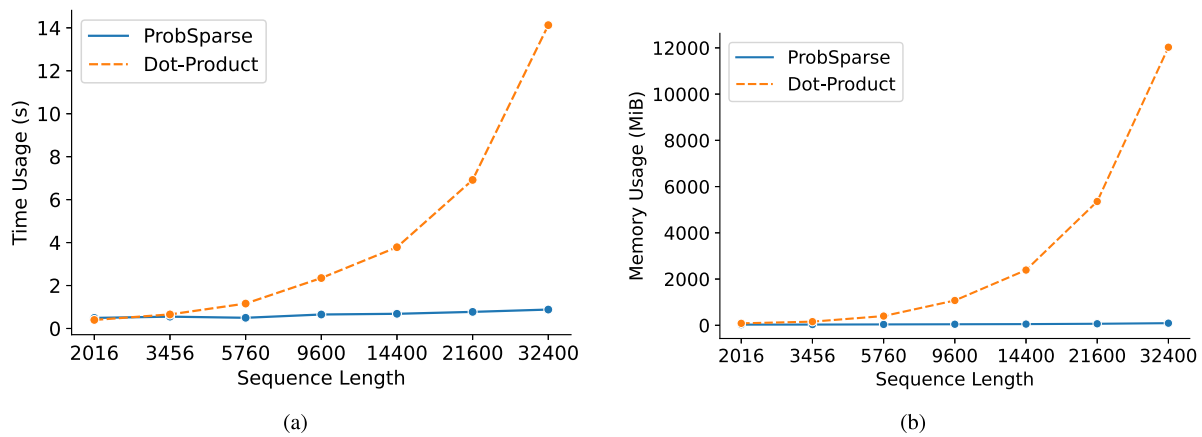


Fig. 10. Growth of computational cost with respect to sequence length for two types of self-attention mechanisms. (a) time usage, (b) memory usage.

Table 7

The impact of distilling on memory usage in BraInf.

Memory usage (MiB) (mean $\pm$ std)	Model block (without distilling)	Model block	Memory usage (MiB) (mean $\pm$ std)
1544.143 $\pm$ 1.736	ProbSparse attention 1 Multilayer perceptron N/A	ProbSparse attention 1 Multilayer perceptron <i>Structural distilling</i>	1491.239 $\pm$ 1.222
1549.785 $\pm$ 1.630	ProbSparse Attention 2 Multilayer perceptron N/A	ProbSparse attention 2 Multilayer perceptron <i>Structural distilling</i>	311.091 $\pm$ 0.574
1544.436 $\pm$ 2.050	ProbSparse attention 3 Multilayer perceptron	ProbSparse attention 3 Multilayer perceptron	59.937 $\pm$ 13.341
Feed-Forward neural network			

Table 8

The impact of distilling on memory usage in Dot-Product models.

Memory usage (MiB) (mean $\pm$ std)	Model block (without distilling)	Model block	Memory usage (MiB) (mean $\pm$ std)
5690.883 $\pm$ 7.951	Dot-Product Attention 1 Multilayer perceptron N/A	Dot-Product Attention 1 Multilayer perceptron <i>Structural distilling</i>	5924.666 $\pm$ 0.279
5703.745 $\pm$ 0.590	Dot-Product Attention 2 Multilayer perceptron N/A	Dot-Product Attention 2 Multilayer perceptron <i>Structural distilling</i>	348.468 $\pm$ 1.477
5704.566 $\pm$ 0.954	Dot-Product Attention 3 Multilayer perceptron	Dot-Product Attention 3 Multilayer perceptron	$\epsilon$
Feed-Forward neural network			

Table 9

Classification performance of different distilling methods.

Task	NC/AD classification			
	ACC (mean $\pm$ std, %)	SEN (mean $\pm$ std, %)	SPE (mean $\pm$ std, %)	PRE (mean $\pm$ std, %)
Original distilling	95.63 $\pm$ 2.50	94.10 $\pm$ 5.54	97.28 $\pm$ 2.10	97.08 $\pm$ 2.20
Structural distilling	97.97 $\pm$ 1.41	97.74 $\pm$ 2.19	98.17 $\pm$ 2.85	98.16 $\pm$ 2.68
Task	NC/MCI Classification			
	ACC (mean $\pm$ std, %)	SEN (mean $\pm$ std, %)	SPE (mean $\pm$ std, %)	PRE (mean $\pm$ std, %)
Original distilling	87.72 $\pm$ 4.54	89.24 $\pm$ 5.08	86.23 $\pm$ 6.09	86.39 $\pm$ 6.39
Structural distilling	91.89 $\pm$ 7.22	90.66 $\pm$ 7.78	93.01 $\pm$ 8.43	92.69 $\pm$ 8.21

distilling remains high. This is also true for the dot-product model shown in Table 8, where  $\epsilon$  denotes very small memory usage.

#### 4.2.3. Classification performance comparisons between structural distilling and original distilling

To validate the effectiveness of our proposed structural distilling in three-dimensional MRI data, we compared its classification performance with the original distilling method described in Section 3.6,

as shown in Table 9. In both the NC/AD and NC/MCI classification tasks, all classification metrics of the proposed structural distilling method were higher than those of the original distilling method. The classification performance improved by 3.26% on average. It was fully demonstrated that the structural distilling method, which considers spatial information and preserves more features, is more applicable to MRI data than the conventional distilling method.

### 4.3. Limitations and future work

Although the model has impressive classification performance, there are two drawbacks at this stage of the study: (1) Fixed patch size. In our implementation, the three-dimensional patch of the feature extraction session had a fixed size of  $5 \times 3 \times 5$ . Since the structural changes in each region caused by brain diseases are not of a fixed size, it is theoretically more appropriate to use a dynamic patch size. In the future, we will attempt to develop a multiscale patch-size self-attention mechanism to make the network dynamic and further enhance the generalization ability of the model. (2) Lack of multiple data modalities. Current studies on brain diseases have used multimodal imaging data [63–65]. Compared with a single MRI modality, multimodal imaging data can provide more information, which can further improve classification performance. Therefore, subsequent studies will try to model multimodal brain data, such as functional MRI [66], Positron Emission Tomography (PET) [67], etc., to achieve better performance.

The self-attention model showed good classification performance on the ADNI dataset. In future works, we hope to further utilize this architecture for lesion analysis. Future studies should aim to identify relevant pathogenic brain regions and extract relevant features from patients' MRI data to provide more reliable scientific evidence for exploring the pathological causes of psychiatric disorders. We validated the excellent classification performance of the model for AD in our experiments. We believe that the proposed model can be easily generalized to other psychiatric disease classification problems, providing a new method for future research on MRI data analysis of psychiatric diseases.

### 5. Conclusion

In this paper, we proposed an efficient model based solely on the self-attention mechanism, called BraInf, to classify MRI data of Alzheimer's disease. The multihead ProbSparse self-attention mechanism used in this study significantly reduces computational complexity, making it possible to apply the self-attention mechanism to high-dimensional MRI data. The structural distilling layer further performs feature down-sampling to retain the key features while reducing the computational cost. The NC/AD and MCI/AD classification accuracies of the proposed architecture on the ADNI dataset were 97.97% and 91.89%, respectively, outperforming other state-of-the-art methods. The experimental results of various ablation studies also illustrate the efficient representation learning capability of the BraInf architecture in brain sMRI data, which provides new ideas and methods for the application of deep learning in the study of brain diseases.

### Acknowledgments

This work was supported in part by the Science and Technology Project in Sichuan, China under grant 2021ZYD0021, 2022NSFSC0530, 2022NSFSC0507, the Sichuan Provincial Program of Traditional Chinese Medicine, China under grant 2021ZD017, and in part by the Fundamental Research Funds for the Central Universities, China, Southwest Minzu University, under grant 2021PTJS23.

### References

- [1] Michael A. DeTure, Dennis W. Dickson, The neuropathological diagnosis of Alzheimer's disease, *Mol. Neurodegener.* 14 (1) (2019) <http://dx.doi.org/10.1186/s13024-019-0333-5>.
- [2] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack Jr., Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, Creighton H. Phelps, The diagnosis of dementia due to Alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease, *Alzheimers Dementia* 7 (3) (2011) 263–269, <http://dx.doi.org/10.1016/j.jalz.2011.03.005>.
- [3] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Carrillo, Bill Thies, Creighton H. Phelps, The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease, *Alzheimers Dementia* 7 (3) (2011) 270–279, <http://dx.doi.org/10.1016/j.jalz.2011.03.008>.
- [4] I.A. Illan, J.M. Gorriz, J. Ramirez, D. Salas-Gonzalez, M.M. Lopez, F. Segovia, R. Chaves, M. Gomez-Rio, C.G. Puntinet, Alzheimer's Dis Neuroimaging Initi, F-18-FDG PET imaging analysis for computer aided Alzheimer's diagnosis, *Inf. Sci.* 181 (4) (2011) 903–916, <http://dx.doi.org/10.1016/j.ins.2010.10.027>.
- [5] Esther E. Bron, Marion Smits, Wiesje M. van der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M. Papma, Rebecca M.E. Steketee, Carolina Mendez Orellana, Rozanna Meijboom, Madalena Pinto, Joana R. Meireles, Carolina Garrett, Antonio J. Bastos-Leite, Ahmed Abdulkadir, Olaf Ronneberger, Nicola Amoroso, Roberto Bellotti, David Cardenas-Pena, Andres M. Alvarez-Meza, Chester V. Dolph, Khan M. Iftekharuddin, Simon F. Eskildsen, Pierrick Coupe, Vladimir S. Fonov, Katja Franke, Christian Gaser, Christian Ledig, Ricardo Guerrero, Tong Tong, Katherine R. Gray, Elaheh Moradi, Jussi Tohka, Alexandre Ruitier, Stanley Durrleman, Alessia Sarica, Giuseppe Di Fatta, Francesco Sensi, Andrea Chincarini, Garry M. Smith, Zhivko V. Stoyanov, Lauge Sorensen, Mads Nielsen, Sabina Tangaro, Paolo Inglese, Christian Wachinger, Martin Reuter, John C. van Swieten, Wiro J. Niessen, Stefan Klein, Alzheimer's Disease Neuroimaging I, Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, *Neuroimage* 111 (2015) 562–579, <http://dx.doi.org/10.1016/j.neuroimage.2015.01.048>.
- [6] Iman Beheshti, Hasan Demirel, Hiroshi Matsuda, Alzheimer's Dis Neuroimaging Initi, Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-alzheimer's conversion from structural magnetic resonance imaging using feature ranking and a genetic algorithm, *Comput. Biol. Med.* 83 (2017) 109–119, <http://dx.doi.org/10.1016/j.combiomed.2017.02.011>.
- [7] Min Young Jung, Young Ok Kim, Woong Yoon, Sung-Pil Joo, Young Jong Woo, Characteristics of brain magnetic resonance images at symptom onset in children with moyamoya disease, *Brain Dev.* 37 (3) (2015) 299–306, <http://dx.doi.org/10.1016/j.braindev.2014.06.008>.
- [8] Tianzi Jiang, Yong Liu, Feng Shi, Ni Shu, Bing Liu, Jiefeng Jiang, Yuan Zhou, Multimodal magnetic resonance imaging for brain disorders: Advances and perspectives, *Brain Imag. Behav.* 2 (4) (2008) 249–257, <http://dx.doi.org/10.1007/s11682-008-9038-z>.
- [9] Zhao Fan, Fanyu Xu, Xuedan Qi, Cai Li, Lili Yao, Classification of Alzheimer's disease based on brain MRI and machine learning, *Neural Comput. Appl.* 32 (7, SI) (2020) 1927–1936, <http://dx.doi.org/10.1007/s00521-019-04495-0>.
- [10] Chunfeng Lian, Mingxia Liu, Jun Zhang, Dinggang Shen, Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI, *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (4) (2020) 880–893, <http://dx.doi.org/10.1109/TPAMI.2018.2889096>.
- [11] S. Sambath Kumar, M. Nandhini, Alzheimers Dis Neuroimaging Initia, Entropy slicing extraction and transfer learning classification for early diagnosis of Alzheimer diseases with sMRI, *ACM Trans. Multimed. Comput. Commun. Appl.* 17 (2) (2021) <http://dx.doi.org/10.1145/3383749>.
- [12] Uttam Khatri, Goo-Rak Kwon, An efficient combination among sMRI, CSF, cognitive score, and andapoe epsilon 4 biomarkers for classification of AD and MCI using extreme learning machine, *Comput. Intell. Neurosci.* 2020 (2020) <http://dx.doi.org/10.1155/2020/8015156>.
- [13] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius, An intelligent system for early recognition of Alzheimer's disease using neuroimaging, *Sensors* 22 (3) (2022) 740.
- [14] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius, Tomas Krilavičius, Analysis of features of Alzheimer's disease: Detection of early stage from functional brain changes in magnetic resonance images using a finetuned ResNet18 network, *Diagnostics* 11 (6) (2021) 1071.
- [15] Imran Razzak, Saeeda Naz, Abida Ashraf, Fahmi Khalifa, Mohamed Reda Bouadjene, Shahid Mumtaz, Multiresolutional ensemble PartialNet for Alzheimer detection using magnetic resonance imaging data, *Int. J. Intell. Syst.* (2022).
- [16] Abida Ashraf, Saeeda Naz, Syed Hamad Shirazi, Imran Razzak, Mukesh Parsad, Deep transfer learning for Alzheimer neurological disorder detection, *Multimedia Tools Appl.* 80 (20) (2021) 30117–30142.
- [17] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, Tsuhan Chen, Recent advances in convolutional neural networks, *Pattern Recognit.* 77 (2018) 354–377, <http://dx.doi.org/10.1016/j.patrec.2017.10.013>.
- [18] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, Muhammad Khurram Khan, Medical image analysis using convolutional neural networks: A review, *J. Med. Syst.* 42 (11) (2018) <http://dx.doi.org/10.1007/s10916-018-1088-1>.
- [19] Bumshik Lee, Waqas Ellahi, Jae Young Choi, Using deep CNN with data permutation scheme for classification of Alzheimer's disease in structural magnetic resonance imaging (sMRD), *IEICE Trans. Inf. Syst.* 102 (7) (2019) 1384–1395.

- [20] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc., 2012, URL <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.
- [21] Yanteng Zhang, Qizhi Teng, Linbo Qing, Yan Liu, Xiaohai He, Lightweight deep residual network for Alzheimer's disease classification using sMRI slices, *J. Intell. Fuzzy Systems* (Preprint) 1–9.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [23] Ahsan Bin Tufail, Yong-Kui Ma, Qiu-Na Zhang, Binary classification of Alzheimer's disease using sMRI imaging modality and deep learning, *J. Digit. Imag.* 33 (5) (2020) 1073–1090.
- [24] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
- [25] Yuma Kinoshita, Hitoshi Kiya, Convolutional neural networks considering local and global features for image enhancement, in: *2019 IEEE International Conference on Image Processing, ICIP, IEEE*, 2019, pp. 2110–2114.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [27] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko, End-to-end object detection with transformers, in: *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.
- [28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
- [30] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, Yunhe Wang, Transformer in transformer, *Adv. Neural Inf. Process. Syst.* 34 (2021).
- [31] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, Hao Ma, Linformer: Self-attention with linear complexity, 2020, arXiv preprint [arXiv:2006.04768](https://arxiv.org/abs/2006.04768).
- [32] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, Wancai Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in: *Proceedings of AAAI*, 2021.
- [33] Clifford R. Jack Jr., Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, Anders M. Dale, Joel P. Felmlee, Jeffrey L. Gunter, Derek L.G. Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colin Studholme, Charles S. DeCarli, Gunnar Krueger, Heidi A. Ward, Gregory J. Metzger, Katherine T. Scott, Richard Mallozzi, Daniel Blezek, Joshua Levy, Josef P. Debbins, Adam S. Fleisher, Marilyn Albert, Robert Green, George Bartzokis, Gary Glover, John Mugler, Michael W. Weiner, ADNI Study, The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods, *J. Magn. Reson. Imag.* 27 (4) (2008) 685–691, [http://dx.doi.org/10.1002/jmri.21049](https://doi.org/10.1002/jmri.21049).
- [34] Christian Gaser, Robert Dahnke, et al., CAT-a computational anatomy toolbox for the analysis of structural MRI data, *Hbm* 2016 (2016) 336–348.
- [35] Andrea Mechelli, Cathy J Price, Karl J Friston, John Ashburner, Voxel-based morphometry of the human brain: Methods and applications, *Curr. Med. Imag.* 1 (2) (2005) 105–113.
- [36] Zhengzhen Li, Jingjing Zhang, Fuqin Wang, Yang Yang, Jie Hu, Qinghui Li, Maoqiang Tian, Tonghuan Li, Bingsheng Huang, Heng Liu, et al., Surface-based morphometry study of the brain in benign childhood epilepsy with centrotemporal spikes, *Ann. Transl. Med.* 8 (18) (2020).
- [37] Yihe Dong, Jean-Baptiste Cordonnier, Andreas Loukas, Attention is not all you need: Pure attention loses rank doubly exponentially with depth, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 2793–2803.
- [38] Alex Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *Physica D* 404 (2020) 132306.
- [39] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to sequence learning with neural networks, *Adv. Neural Inf. Process. Syst.* 27 (2014).
- [40] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint [arXiv:1409.0473](https://arxiv.org/abs/1409.0473).
- [41] Rewon Child, Scott Gray, Alec Radford, Ilya Sutskever, Generating long sequences with sparse transformers, 2019, <https://arxiv.org/abs/1904.10509>, URL <https://arxiv.org/abs/1904.10509>.
- [42] Nikita Kitaev, Lukasz Kaiser, Anselm Levskaya, Reformer: The efficient transformer, in: *International Conference on Learning Representations*, 2020, URL <https://openreview.net/forum?id=rrkgNkKHtVb>.
- [43] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, Ruslan Salakhutdinov, Transformer-XL: Attentive language models beyond a fixed-length context, 2019, CoRR, [abs/1901.02860](https://arxiv.org/abs/1901.02860) [arXiv:1901.02860](https://arxiv.org/abs/1901.02860).
- [44] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, Ruslan Salakhutdinov, Transformer dissection: A unified understanding of transformer's attention via the lens of kernel, 2019, arXiv preprint [arXiv:1908.11775](https://arxiv.org/abs/1908.11775).
- [45] Djork-Arné Clevert, Thomas Unterthiner, Sepp Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), 2015, arXiv preprint [arXiv:1511.07289](https://arxiv.org/abs/1511.07289).
- [46] Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras, Spam filtering with naive Bayes - which naive Bayes? in: CEAS, 2006.
- [47] David W. Hosmer Jr., Stanley Lemeshow, Rodney X. Sturdivant, *Applied Logistic Regression*, Vol. 398, John Wiley & Sons, 2013.
- [48] Johannes Fürnkranz, Decision tree, in: Claude Sammut, Geoffrey I. Webb (Eds.), *Encyclopedia of Machine Learning*, Springer US, Boston, MA, 2010, pp. 263–267, [http://dx.doi.org/10.1007/978-0-387-30164-8\\_204](https://doi.org/10.1007/978-0-387-30164-8_204).
- [49] Abraham J. Wyner, Matthew Olson, Justin Bleich, David Mease, Explaining the success of AdaBoost and random forests as interpolating classifiers, *J. Mach. Learn. Res.* 18 (48) (2017) 1–33, URL <http://jmlr.org/papers/v18/15-240.html>.
- [50] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [51] Janez Demšar, Statistical comparisons of classifiers over multiple data sets, *J. Mach. Learn. Res.* 7 (2006) 1–30.
- [52] Alessio Benavoli, Giorgio Corani, Francesca Mangili, Should we really use post-hoc tests based on mean-ranks? *J. Mach. Learn. Res.* 17 (1) (2016) 152–161.
- [53] Jinwang Feng, Shao-Wu Zhang, Luonan Chen, Jie Xia, Alzheimer's Dis Neuroimaging Init, Alzheimer's disease classification using features extracted from nonsubsampling contourlet subband-based individual networks, *Neurocomputing* 421 (2021) 260–272.
- [54] Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, Mingqing Xu, Alzheimer's Dis Neuroimaging Init, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, *Neuroimage* 208 (2020) [http://dx.doi.org/10.1016/j.neuroimage.2019.116459](https://doi.org/10.1016/j.neuroimage.2019.116459).
- [55] Wanyun Lin, Weiming Lin, Gang Chen, Hejun Zhang, Qinquan Gao, Yechong Huang, Tong Tong, Min Du, Alzheimer's Disease Neuroimaging, Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer's disease, *Front. Neurosci.* 15 (2021) [http://dx.doi.org/10.3389/fnins.2021.646013](https://doi.org/10.3389/fnins.2021.646013).
- [56] Zhenyuan Ning, Qing Xiao, Qianjin Feng, Wufan Chen, Yu Zhang, Relation-induced multi-modal shared representation learning for Alzheimer's disease diagnosis, *IEEE Trans. Med. Imag.* 40 (6) (2021) 1632–1645, [http://dx.doi.org/10.1109/TMI.2021.3063150](https://doi.org/10.1109/TMI.2021.3063150).
- [57] R. Divya, R. Shantha Selva Kumari, Alzheimer's Dis Neuroimaging Initia, Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification, *Neural Comput. Appl.* 33 (14, SI) (2021) 8435–8444, [http://dx.doi.org/10.1007/s00521-020-05596-x](https://doi.org/10.1007/s00521-020-05596-x).
- [58] Wenjie Kang, Lan Lin, Baiwen Zhang, Xiaoqi Shen, Shuicai Wu, Multi-model and multi-slice ensemble learning architecture based on 2D convolutional neural networks for Alzheimer's disease diagnosis, *Comput. Biol. Med.* 136 (2021) 104678, [http://dx.doi.org/10.1016/j.compbiomed.2021.104678](https://doi.org/10.1016/j.compbiomed.2021.104678), URL <https://www.sciencedirect.com/science/article/pii/S0010482521004728>.
- [59] Mohammed Abdelaziz, Tianfu Wang, Ahmed Elazab, Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks, *J. Biomed. Inf.* 121 (2021) [http://dx.doi.org/10.1016/j.jbi.2021.103863](https://doi.org/10.1016/j.jbi.2021.103863).
- [60] Jayanthi Venkatraman Shanmugam, Baskar Duraisamy, Blessy Chittatukarakaran Simon, Preethi Bhaskaran, Alzheimer's disease classification using pre-trained deep networks, *Biomed. Signal Process. Control* 71 (2022) 103217, [http://dx.doi.org/10.1016/j.bspc.2021.103217](https://doi.org/10.1016/j.bspc.2021.103217), URL <https://www.sciencedirect.com/science/article/pii/S1746809421008144>.
- [61] Nitika Goenka, Shamik Tiwari, AlzyNet: A volumetric convolutional neural network for multiclass classification of Alzheimer's disease through multiple neuroimaging computational approaches, *Biomed. Signal Process. Control* 74 (2022) 103500, [http://dx.doi.org/10.1016/j.bspc.2022.103500](https://doi.org/10.1016/j.bspc.2022.103500), URL <https://www.sciencedirect.com/science/article/pii/S1746809422000222>.
- [62] Jianguang Li, Ying Wei, Chuyuan Wang, Qian Hu, Yue Liu, Long Xu, 3-D CNN-based multichannel contrastive learning for Alzheimer's disease automatic diagnosis, *IEEE Trans. Instrum. Meas.* 71 (2022) 1–11, [http://dx.doi.org/10.1109/TIM.2022.3162265](https://doi.org/10.1109/TIM.2022.3162265).
- [63] Jingcong Li, Zhu Liang Yu, Zhenghui Gu, Hui Liu, Yuanqing Li, MMAN: Multi-modality aggregation network for brain segmentation from MR images, *Neurocomputing* 358 (2019) 10–19, [http://dx.doi.org/10.1016/j.neucom.2019.05.025](https://doi.org/10.1016/j.neucom.2019.05.025).
- [64] Nur Suriza Syazwany, Ju-Hyeon Nam, Sang-Chul Lee, MM-BiFPN: Multi-modality fusion network with Bi-FPN for MRI brain tumor segmentation, *IEEE Access* 9 (2021) 160708–160720, [http://dx.doi.org/10.1109/ACCESS.2021.3132050](https://doi.org/10.1109/ACCESS.2021.3132050).
- [65] Lei Du, Kefei Liu, Xiaohui Yao, Shannon L. Risacher, Junwei Han, Andrew J. Saykin, Lei Guo, Li Shen, Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics, *IEEE-ACM Trans. Comput. Biol. Bioinform.* 18 (1) (2021) 227–239, [http://dx.doi.org/10.1109/TCBB.2019.2947428](https://doi.org/10.1109/TCBB.2019.2947428).
- [66] Gary H. Glover, Overview of functional magnetic resonance imaging, *Neurosurg. Clin.* 22 (2) (2011) 133–139.
- [67] Arvind K. Shukla, Utham Kumar, Positron emission tomography: An overview, *J. Med. Phys./Assoc. Med. Physicists India* 31 (1) (2006) 13.