Computer Methods and Programs in Biomedicine 229 (2023) 107291

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction

Zhentao Hu[a], Zheng Wang[a,*], Yong Jin[a], Wei Hou[b]

[a] School of Artificial Intelligence, Henan University, Zhengzhou, 450046, China
[b] College of Computer and Information Engineering, Henan University, Kaifeng, 475004, China

ARTICLE INFO

Article history:
Received 6 April 2022
Revised 22 November 2022
Accepted 28 November 2022

Keywords:
Alzheimer's disease (AD)
Deep learning
Longitudinal study
Convolutional neural network (CNN)
Transformer

ABSTRACT

Background and objective: Mild cognitive impairment (MCI) is a transitional state between normal aging and Alzheimer's disease (AD), and accurately predicting the progression trend of MCI is critical to the early prevention and treatment of AD. Brain structural magnetic resonance imaging (sMRI), as one of the most important biomarkers for the diagnosis of AD, has been applied in various deep learning models. However, due to the inherent disadvantage of deep learning in dealing with longitudinal medical image data, few applications of deep learning for longitudinal analysis of MCI, and the majority of existing deep learning algorithms for MCI progress prediction rely on the analysis of the sMRI images collected at a single time-point, ignoring the progressive nature of the disorder. Methods: In this work, we propose a VGG-TSwinformer model based on convolutional neural network (CNN) and Transformer for short-term longitudinal study of MCI. In this model, VGG-16 based CNN is used to extract low-level spatial features of longitudinal sMRI images and map these low-level features to high-level feature representations, sliding-window attention is used for fine-grained fusion of spatially adjacent feature representations, and gradually fuses distant spatial feature representations through the superposition of attention windows of different sizes, temporal attention is used to measure the evolution of this feature representations as a result of disease progression. Results: We validated our model on the ADNI dataset. For the classification task of sMCI vs pMCI, accuracy, sensitivity, specificity and AUC reached 77.2%, 79.97%, 71.59% and 0.8153 respectively. Compared with other cross-sectional studies also applied to sMRI, the proposed model achieved better results in terms of accuracy, sensitivity, and AUC. Conclusion: The proposed VGG-TSwinformer is a deep learning model for short-term longitudinal study of MCI, which can build brain atrophy progression model from longitudinal sMRI images, and improve diagnostic efficiency compared to algorithms using only cross-sectional sMRI images.

1. Introduction

The increasingly serious population aging has brought about various senile diseases, among which Alzheimer's disease (AD) is one of the most intractable diseases. According to 2022 Alzheimer's Association report, there are currently 6.5 million Americans age 65 and older living with AD. Without medical breakthroughs to prevent, slow or treat AD, that number could grow to 13.8 million by 2060 [1]. AD is a degenerative and irreversible brain disease, with the progression of the disease, more and more brain neurons stop functioning, lose connection, and even die. In addition, there is a corporal change in the common AD-related variation of anatomical brain structures such as the enlargement of ventricles, shrinkage of the hippocampus shape, change in the cortical thickness, and other cerebral areas containing white matter and gray matter brain tissue as well as cerebrospinal fluid [2].

The mild cognitive impairment (MCI) is a transitional state between the normal aging and AD, and about 44% of MCI patients will convert to AD within 3 years [3]. As shown in Fig. 1, for potential AD patients, when they progress from cognitively normal (CN) to MCI, will rapidly progress into AD [4]. The DSM-5 criteria [5] divided MCI patients who transformed into AD in the next 3 years into the progressive MCI (pMCI) group, and MCI patients who did not transform into AD in the next 3 years into the stable MCI (sMCI) group. People with MCI experience cognitive decline and the structure of their brains gradually shrinks. Jack et al. [6] measured the hippocampal and entorhinal cortex, whole brain and ventricular volumes of CN, MCI and AD patients, and calculated the rate of brain atrophy in 1–5 years, the results shown that

[*] Corresponding author.
E-mail address: wangzheng@henu.edu.cn (Z. Wang).

https://doi.org/10.1016/j.cmpb.2022.107291
0169-2607/© 2022 Elsevier B.V. All rights reserved.

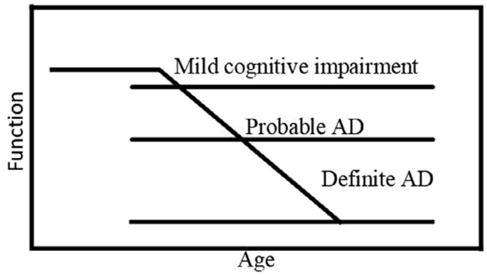

Fig. 1. Theoretic progression of cognitive function from normal, through MCI, to probable and definite AD in potential AD patients.

the annual shrinkage rate of MCI with a tendency to convert to AD was the fastest, followed by stable MCI, and the slowest was CN. Although MCI patients have impairments in language, memory, thinking ability, etc., their symptoms are not as severe as AD, so MCI is often mistaken as a manifestation of aging and not taken seriously [7], however, when MCI is completely transformed into AD, there are no effective drugs and treatments that can cure AD so far. Distinguishing pMCI and sMCI plays an important role in the early diagnosis of AD, which can help clinicians to take effective interventions for the progression of the disease [8]. However, the classification of pMCI and sMCI is more challenging than that of AD and CN because of the more subtle differences in cognition and brain structure between pMCI and sMCI patients [9].

There are various methods for the diagnosis of AD, including cerebrospinal fluid puncture, electroencephalography, genetic testing, neuropsychological examination, neuroimaging examination, etc [10]. Among them, structural magnetic resonance imaging (sMRI) can intuitively reflect the pathological changes of patients, and because of its non-invasive advantages, it is more easily accepted by patients and their families. In addition, sMRI can detect the time course of brain atrophy in patients with AD and may serve as a surrogate marker for pathological changes in people with suspected AD [11]. More recently, with the rise of computer-aided diagnosis, a number of algorithms have been developed to predict MCI based on sMRI images. Basically, these algorithms use sMRI images acquired at a single time-point to predicts diagnosis labels, ignoring the progressive nature of the disorder. Considering that the incubation period of AD can last for decades, simple interpretation of brain histomorphology changes from a single cross-section cannot clearly reveal the evolution of the disease, and follow-up observations at multiple time points may be of clinical significance [12]. However, due to the inherent disadvantage of deep learning in dealing with longitudinal medical image data, there are few applications of deep learning for longitudinal analysis of MCI.

This paper proposes a deep learning model called VGG-TSwinformer based on longitudinal sMRI images to build the pattern of changes in brain structure over time in MCI patients. In addition, a fixed time interval of 2 years is introduced into the longitudinal sMRI images, which helps establish a timeline of the disease progression. In other words, this study considers the whole-brain morphology of MCI patients at two time points, and models the whole-brain morphology at these two time points by VGG-TSwinformer, so that the change pattern of the patient's whole-brain morphology could be obtained. By fixing the interval between these two time points, the evolution rate of the disease could be introduced, so that the future clinical status of MCI patients could be grasped. We aim to propose a fully automated workflow for short-term longitudinal studies of MCI, including MRI preprocessing, MRI feature extraction, MRI feature fusion and classification, that can automatically identify brain changes without manual labeling by experts, thereby achieve good generalization ability. Our contributions are as follows:

1. In this study, a deep learning model VGG-TSwinformer based on longitudinal sMRI images is proposed to establish a model of brain atrophy progression in MCI patients. In addition, we introduce a fixed time interval of 2 years to the longitudinal sMRI images, which helps to establish a timeline of the disease progression.
2. A new sliding-window attention mechanism is proposed for fine-grained fusion of adjacent spatial features of sMRI image, and gradually fuse long-distance spatial features through the superposition of attention windows of different sizes.
3. Temporal attention is used to establish patterns of brain structural changes due to disease progression, which are reflected in longitudinal sMRI images.

The remainder of the paper is structured as follows: Section II summarizes some of the related works and gives a brief introduction to the VGG-TSwinformer model. Section III introduces the architecture of VGG-TSwinformer model in detail. Section IV introduces the experimental results and details. Finally, we summarize our work and make a prospect for future work in Section V.

2. Related works

Deep learning is a powerful technique, state-of-the-art for many machine learning (ML) problems. Recently, deep learning-based assisted AD diagnosis has achieved great success, even surpassing manual diagnosis. AD diagnostic methods can be divided into two groups according to the task they aim: progression models, which seek to quantify the evolution of the disease, such as pMCI vs sMCI, and classification models, which predicts diagnosis labels of patients, such as AD vs CN [13].

For progression models, Suk et al. [14] proposed a novel framework that combines the two conceptually different methods of sparse regression and deep learning for AD and MCI diagnosis and prediction, where CNNs were used as weak learners to extract low-level features, and multiple sparse regression models were used for final prediction, and 74% pMCI vs sMCI classification accuracy achieved on ADNI dataset. Basaia et al. [15] designed a architecture based on 3D CNN, and the classification accuracy of pMCI vs sMCI on ADNI dataset reached 75.1%. Li and Liu [16] proposed a network combining 3D CNN and RNN. 3D CNN extracted the left and right hippocampal features of a MRI single image, and the RNN established the relationship between the left and right hippocampal features. In the prediction tasks of pMCI vs sMCI, the classification accuracy reached 72.5%. Oh et al. [17] constructed an unsupervised learning model based on convolutional autoencoder (CAE) to classify AD and CN, and transferred the model to the more complex classification task of pMCI vs sMCI, achieving an accuracy of 73.95%. Recently, Altay et al. [18] has proposed three deep learning networks: 3D CNN Model, 3D Recurrent Visual Attention Model and Attention Transformer Model for predicting the preclinical stage of AD, among them, the Attention Transformer Model used CNN to extract slices features of 3D sMRI image, and used Transformer for feature fusion between slices. In fact, both the prediction of the preclinical stage of AD and the prediction of pMCI/sMCI are progressive predictions. In conclusion, Suk et al. [14], Basaia et al. [15], Li and Liu [16], Oh et al. [17], Altay et al. [18] make full use of the cross-sectional sMRI images to make predictions, however, like most classification models, may have a good performance on the current state of the disease predictions, but there are inherent drawbacks to progressive predictions due to the inability to consider longitudinal linkages.

Longitudinal studies of MCI based on sMRI image data can be divided into long-term longitudinal studies and short-term longitudinal studies according to the follow-up time. Long-term longitudinal studies can clearly track the changes in brain morphology of MCI patients over time. But in deep learning, long-term longitudinal studies mean processing long sequences of 3D medical image data, which will lead to high-dimensional data volume, and using small datasets to train high-dimensional data has the risk of overfitting. Aghili et al. [19] manually extracted MRI cortical thickness, volume, shape of hippocampal and voxel-wise tissue features at each time point, and used RNN to establish a temporal relationship between the extracted features. However, since the feature extraction and classifier model were independent, the extracted features may not capture the full characteristics of brain abnormalities related to AD. Compared with long-term longitudinal studies, the MRI datasets of short-term longitudinal studies are relatively sufficient, however, the changes in brain morphogenesis in the short-term of MCI patients are not obvious enough. In addition, MRI datasets for short-term longitudinal studies are still far less large than those for cross-sectional studies, so capturing such subtle brain structural changes to predict MCI progression in limited iterative training becomes a challenge for short-term longitudinal studies. Cui and Liu [20] and Cui et al. [21] used 3D CNN and multilayer perceptron (MLP) to extract brain morphological features from MRI, respectively, and RNN fused these features to extract longitudinal change information. However, feature representation of the entire MRI may result in weakened or even lost local spatial features. In addition, the output of RNN at next time point must be based on the output of previous time point, and cannot be processed in parallel.

Transformer's success in computer vision may advance the application of deep learning in short-term longitudinal disease studies. Transformer, which was first applied in the field of natural language processing [22], is a deep neural network based on self-attention. Transformer utilizes self-attention mechanism to capture long-range dependencies, enabling the model to focus on all elements in input sequence, resulting in better performance. Recently, Transformer has made breakthroughs in computer vision. For computer vision tasks, a large number of Transformer-based methods have been proposed, such as DETR [23] for object detection, SETR [24] for semantic segmentation, ViT [25] and DeiT [26] for image recognition. Taking Swin Transformer [27] as vision backbone, the authors achieved state-of-the-art performance on image classification, object detection and semantic segmentation. In Swin Transformer, self-attention computations are restricted to non-overlapping local windows, while also allowing cross-window connections. The window attention enhances the local feature extraction of the image, and global image features can be gradually extracted through the superposition of attention windows. As an end-to-end image recognition method, CNN has dominated the computer vision field in the past few years: AlexNet [28], VGGNet [29], GoogLeNet [30], and ResNet [31]. Compared with Transformer, CNN uses local receptive fields, shared weights and spatial downsampling [32], which has the advantages of learning and summarizing low-level features [33].

Considering the advantages of CNN in summarizing low-level features and Transformer in processing remote features, this paper proposes a VGG-TSwinformer model based on CNN and Tranformer, the overall architecture of the model is shown in Fig. 5. We use two sMRI images acquired at time points T1 and T2 as the longitudinal data of MCI patient, denoted as T1 sMRI image and T2 sMRI image. T1 sMRI image and T2 sMRI image are sliced to obtain two 2D slice series: slice series T1 and slice series T2, slice series T1 and slice series T2 are taken as a sample belonging to one patient. VGG-16 based CNN is used to extract the low-level spatial features of slices and encode the slices as high-level feature representation tokens, and each token corresponds to a 2D slice, finally get two token series: token series T1 and token series T2. Inspired by Swin Transformer [27], we propose an sliding-window attention mechanism to fuse local spatial information of token series T1 and token series T2, respectively, and the global spatial features are gradually integrated with the superposition of windows. Our proposed temporal attention can perform feature fusion between tokens belonging to token series T1 and token series T2 to extract the local atrophy progression pattern of patients' brain structures in longitudinal time T1 to T2. Alternate execution of temporal attention and spatial attention establishes a spatial connection between the change information generated by the local features, and finally obtains the global feature change information, which can be used to predict the progress of MCI.

3. Methods

This section describes the methodology that relevant to this work, including VGG-16 and Transformer, and then introduces the whole VGG-TSwinformer model in detail.

3.1. VGG-16 based CNN

The structure of the brain is relatively complex and is roughly divided into 6 parts, each of which contains multiple brain regions. VGG-16 reduces the size of the convolution kernel and increases the number of convolution kernels on the basis of AlexNet [28], so it is more suitable for brain image processing and can extract more abstract features.

In this paper, a VGG-16 based CNN is used to extract the spatial features of the slices in slice series T1 and slice series T2 of each sample and map them to high-level feature representation tokens. The VGG-16 based CNN architecture is shown in Fig. 2. VGG-16 based CNN uses 13 convolutional layers and 5 pooling layers of VGG-16, and adds a convolutional layer at the head for channel expansion of the slice, and a convolutional layer at the tail for mapping from feature map to token. The slice is dimensioned to $112 \times 112 \times 3$ after dimension expansion. After a series of convolution, activation, pooling operations in VGG-16, the size of the output feature map of is $3 \times 3 \times 512$. Finally, a convolution layer with kernel-size of 3×3, input-channel of 512, output-channel of 256 performs the mapping from feature map to token. The standard convolution calculation process formula of the feature map I to the feature map O is given by

$$O_j = ReLU\left(\sum_{i \in M_j} I_i * k_{ij} + b_j\right) \tag{1}$$

where O_j is j channel of feature map O, I_i is i channel of feature map I, M_j is the set of channels in feature map I, k_{ij} represents the convolutional kernel corresponding to I_i and O_j, b_j represents the bias offset of O_j after convolution, $*$ represents the convolutional calculation. ReLU is an activation function, which turns all negative inputs into 0, while positive values remain unchanged. For input t, ReLU is expressed as

$$ReLU(t) = \begin{cases} t, & t > 0 \\ 0, & t \leq 0 \end{cases} \tag{2}$$

Assuming that the feature map P is obtained by max pooling the feature map O, the feature map P can be given by

$$\begin{aligned} P_j(h,w) =& \omega_j \bullet \max_{k_h,k_w} O_j(hs+k_h, ws+k_w) + b_j, \\ & k_h \in [1,2,\ldots,K], k_w \in [1,2,\ldots,K] \end{aligned} \tag{3}$$

where P_j, O_j is the j channel of the feature map P and feature map O respectively, h, w is the row number and column number of P_j respectively, s is the stride of the pooling window, K is the

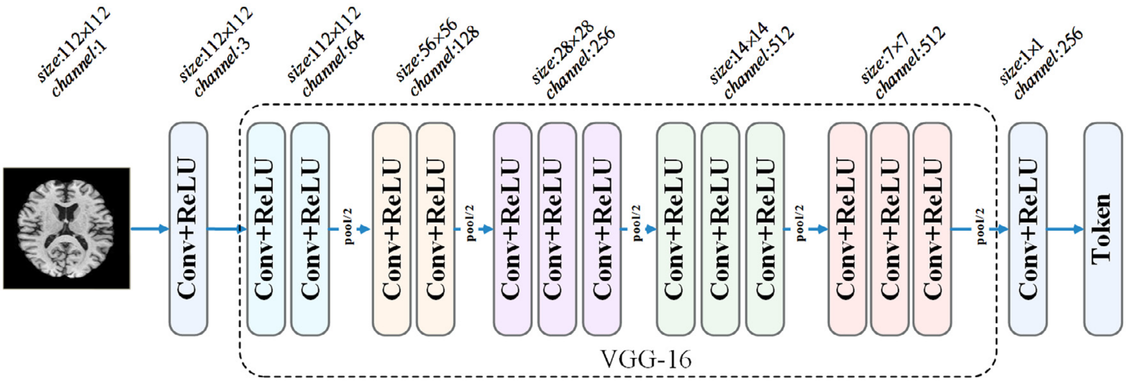

Fig. 2. VGG-16 based CNN architecture.

size of the pooling window, ω_j and b_j represent the weight coefficients and bias offset parameters respectively, and different input feature maps correspond to different weight coefficients and bias offset parameters. The convolution operation in VGG-16 extracts the lowest-level features of the input image. The features extracted by each convolution kernel are different, and the number of local features extracted by the convolution layer groups in turn is 64, 128, 256 and 512. The last convolutional layer filters out 256 of the 512 features as the feature representation for this slice.

Convolution mapping of feature map to token is performed on N slices in slice series T1 and slice series T2 respectively, and get two token series: token series T1 and token series T2, finally each sample has a total of 2N tokens. We use sinusoidal position encoding [22] to embed the temporal position of a pair of tokens in the same position of token series T1 and token series T2, and embed the spatial position within the series for token series T1 and token series T2 respectively. The 2N tokens are sent to the first temporal attention block after spatial and temporal position embedding,

3.2. Transformer

In natural language processing (NLP), Transformer consists of two parts: Encoder and Decoder. The Encoder maps the input (language sequence) into hidden layers, and the decoder maps the hidden layers to natural language sequences. Since remapping is not required in image processing, we follow ViT [25] and only use the Encoder part in Transformer (hereinafter referred to as Transformer).

3.2.1. Self-attention (SA)

SA is a feature fusion method to capture the internal correlation of features. SA reduces the weight of unimportant information while tilting the weight towards useful information. The calculation process of SA is as follows: first, input tokens are linearly mapped to generate a set of query matrix Q, key matrix K and value matrix V, and the weight coefficients are calculated by Q and K,and then use the obtained weight coefficients to compute the weighted sum of the vectors in K, that is, the weighted fusion between the tokens. As shown in Fig. 3, the SA is calculated as

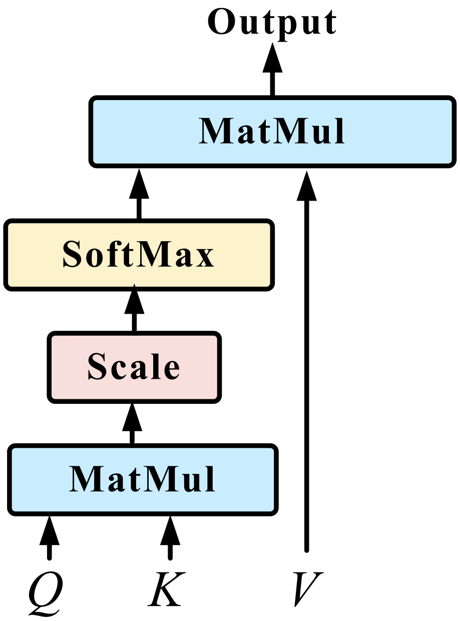

Fig. 3. Self-attention calculation process.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{4}$$

where d_k is the dimension of key vector K, $\sqrt{d_K}$ provides an appropriate normalization to make the gradient more stable.

3.2.2. Multi-head self-attention (MSA)

MSA means that input tokens generate h groups (called h heads) of query matrix, key matrix and value matrix, and perform SA operations in each group of query matrix, key matrix and value matrix. Finally, concatenate the outputs of h heads. As shown in Fig. 4, the MSA can be given by

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

$$MSA(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{6}$$

where the projections W_i^Q, W_i^K, W_i^V and W^O are trainable parameter matrices, h is the number of MSA heads.

Combining the above, a complete Transformer block can be given by

$$t(x) = MSA(LN(x)) + x \tag{7}$$

$$Transformer(x) = MLP(LN(t(x))) + t(x) \tag{8}$$

where x is the input tokens and LN is the layer normalization [34], which is applied before MSA and MLP. The MLP consists of two fully connected layers and residual links, and the activation function GELU [35] is used in the first fully connected layer.

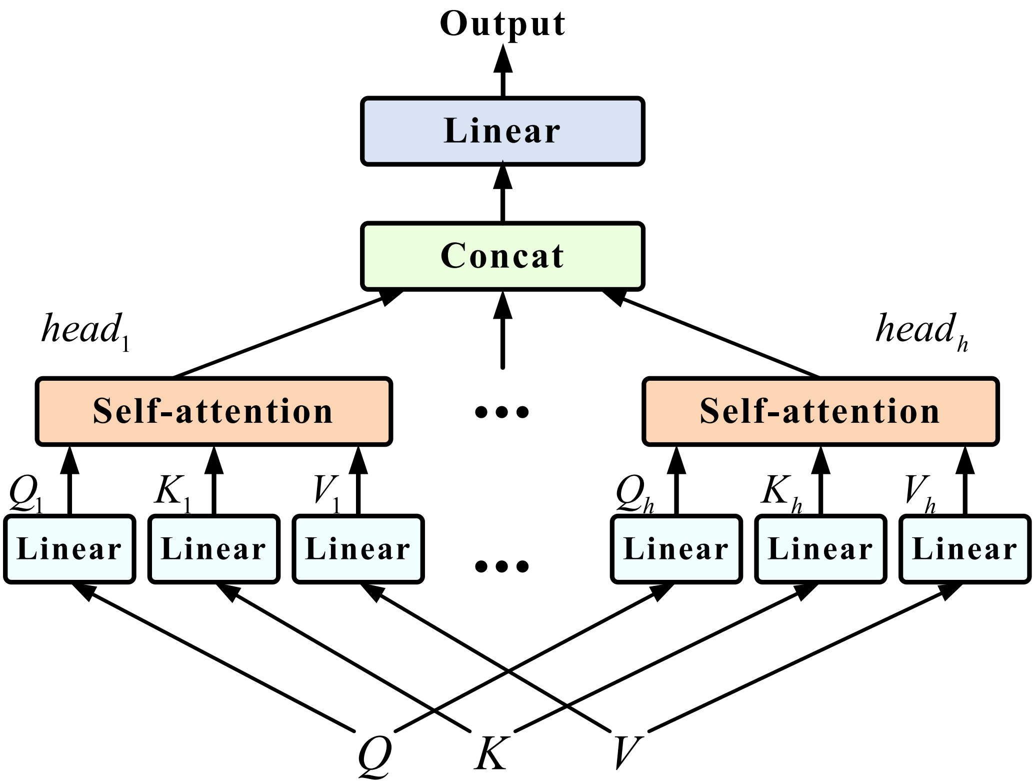

Fig. 4. Multi-head self-attention calculation process.

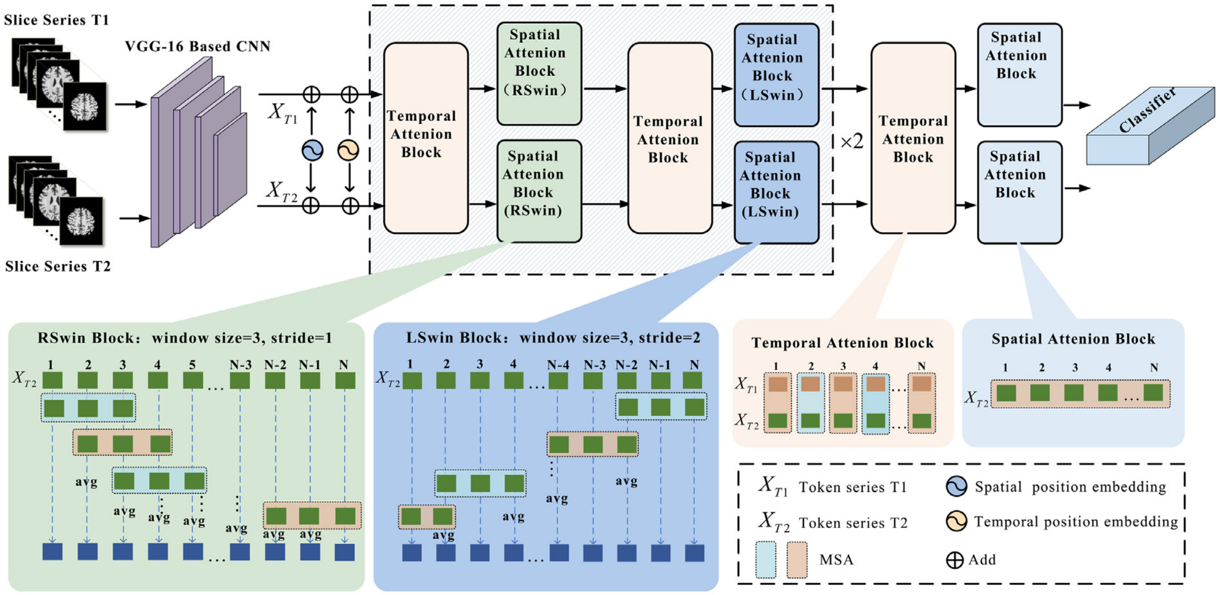

Fig. 5. VGG-TSwinformer model architecture.

3.3. VGG-TSwinformer model

The architecture of the VGG-TSwinformer proposed in this paper is shown in Fig. 5. VGG-16 based CNN maps each slice to a high-level feature representation token. Finally, token series T1 and token series T2 representing slice series T1 and slice series T2 respectively are sent to the first temporal attention block. In VGG-TSwinformer, token series T1 and token series T2 each have 10 attention blocks, of which token series T1 and token series T2 each have 5 spatial attention blocks, and share 5 temporal attention blocks. In order to better integrate local features and try to avoid dividing the same redundant tokens, the first four spatial attention blocks of token series T1 and token series T2 are designed alternately into right-sliding window (RSwin) attention block and left-sliding window (LSwin) attention block.

For short-term longitudinal studies of MCI, we need to focus on changes in brain morphogenesis over time, which are subtle and not easily detected. Therefore, as shown in Fig. 5, in the temporal attention block, we do MSA to the tokens at the corresponding positions of token series T1 and token series T2 to enhance the extraction of local longitudinal features, that is, we perform feature fusion on the axial slices at the corresponding positions of the T1 sMRI image and T2 sMRI image. In addition, our proposed siding-window attention mechanism enables indirect feature fusion of tokens within the corresponding window range of token series T1 and token series T2, so that the feature fusion of T1 sMRI image and T2 sMRI image extends from the corresponding 2D slice to the local 3D space. Generally, features with closer spatial distances tend to have stronger correlations than features with longer spatial distances. For small sMRI datasets, the proposed siding-window attention and temporal attention can make it easier for the model to notice the changes of local features in limited iterative training, so as to use this change information for prediction.

Suppose the lth block is temporal attention block, the $l+1$th block is RSwin (or LSwin) attention block, and the output of the $l-1$th block is $X_{T1}^{l-1}$and X_{T2}^{l-1}. X_{T1}^{l-1} and X_{T2}^{l-1} can be expressed as

$$\begin{cases} X_{T1}^{l-1} = (X_{(T1,1)}^{l-1}; \cdots ; X_{(T1,N)}^{l-1}) \\ X_{T2}^{l-1} = (X_{(T2,1)}^{l-1}; \cdots ; X_{(T2,N)}^{l-1}) \end{cases} \tag{9}$$

where X_{T1}^{l-1}, X_{T2}^{l-1} are token series T1 and token series T2 after the $l-1$th block, respectively, and $X_{(T1,i)}^{l-1}$, $X_{(T2,i)}^{l-1}$ are the ith token of X_{T1}^{l-1} and X_{T2}^{l-1}, respectively.

The lth temporal attention block is to perform MSA within the two tokens in the corresponding positions of the token series T1 and token series T2. Suppose the output of the lth block is $X_{T1}^{l}$and X_{T2}^{l}, X_{T1}^{l} and X_{T2}^{l} can be expressed as

$$\begin{cases} X_{T1}^{l} = (X_{(T1,1)}^{l}; \cdots ; X_{(T1,N)}^{l}) \\ X_{T2}^{l} = (X_{(T2,1)}^{l}; \cdots ; X_{(T2,N)}^{l}) \end{cases} \tag{10}$$

and

$$(X_{(T1,i)}^{l}, X_{(T2,i)}^{l}) = MSA(X_{(T1,i)}^{l-1}; X_{(T2,i)}^{l-1}), i = 1, \ldots, N \tag{11}$$

where X_{T1}^{l}, X_{T2}^{l} are token series T1 and token series T2 after the lth temporal attention block, respectively, and $X_{(T1,i)}^{l}$, $X_{(T2,i)}^{l}$ are the ith tokens of X_{T1}^{l} and X_{T2}^{l}, respectively.

Then perform the $l+1$th RSwin (or LSwin) attention block on X_{T1}^{l} and X_{T2}^{l} respectively. As shown in Fig. 5, as a demonstration, in the RSwin block, the division of windows is from left to right. We set the window size to 3 and the window sliding stride to 1. The stride determines the number of overlapping tokens between the two windows before and after. We average the overlapping tokens as the output of this block at this position, thus preserving the boundary information of adjacent windows. In the LSwin block in Fig. 5, the division of the window is from right to left. We set the window size to 3 and the window sliding stride to 2. It can be seen that there are still redundant tokens after the complete window is divided, and the window continues to slide to the left to generate a new incomplete window and perform MSA within the window.

Suppose X_{T1}^{l+1} and X_{T2}^{l+1} are obtained by X_{T1}^{l} and X_{T2}^{l} after RSwin (or LSwin) attention block, respectively. Algorithms 1 and 2 show the calculation process of RSwin and LSwin block of X_{T2}^{l+1} respectively, where W_i represents the output of ith attention window, w is the size of the attention window, N is the number of tokens contained in the token series T1 or token series T2, num is the total number of full windows in this RSwin (or LSwin) attention block, s is the stride of window sliding and s < w, $ceil()$ is the upward integer function, $M \in \mathbb{R}^{(num+1)\times N\times C}$ is used to store the calculation results of each window. For X_{T1}^{l+1}, the calculation process is the same as the calculation process of $X_{(T2)}^{l+1}$.

As shown in Fig. 5, in the last spatial attention block of token series T2, MSA is done for all tokens in token series T2. For token series T1, the calculation process is the same as the calculation process of token series T2. Configuration details of four sliding-window attention blocks of token series T1 and token series T2 are shown in Table 1. The output tokens of the last spatial attention block of token series T1 and token series T2 are averaged and sent to the classifier to get the final prediction.

Algorithm 1 RSwin block calculation process.

```
Input: $X_{T2}^{l} = (X_{(T2,1)}^{l}; \ldots; X_{(T2,N)}^{l})$
Output: $X_{T2}^{l+1}$
$num = ceil(\frac{N-w+1}{s})$
$M \leftarrow zero(num+1, N, C)$
$i \leftarrow 1$ // window number
while ($i \le num$) do
    $W_i \leftarrow MSA(X_{(T2,(i-1)s+1)}^{l}; \ldots; X_{(T2,(i-1)s+w)}^{l})$
    $M[i-1, (i-1)s : (i-1)s + w - 1, :] \leftarrow W_i$
    $i \leftarrow i + 1$
end while
if $((num-1)s + w ! = N)$ then //complete window are divided and redundant tokens exist
    $W_i \leftarrow MSA(X_{(T2,(i-1)s+1)}^{l}, \ldots, X_{(T2,N)}^{l})$
    $M[i-1, num * s : N-1, :] \leftarrow W_i$
end if
$count \leftarrow (M[:, :, C] ! = zero(C)).sum(axis = 0)$ // count the number of overlapping tokens for each location
$t \leftarrow M.sum(axis = 0)/count$ // average to get the final output
$X_{T2}^{l+1} \leftarrow MLP(LN(t)) + t$
return $X_{T2}^{l+1}$
```

Algorithm 2 LSwin block calculation process.

```
Input: $X_{T2}^{l} = (X_{(T2,1)}^{l}; \ldots; X_{(T2,N)}^{l})$
Output: $X_{T2}^{l+1}$
$num = ceil(\frac{N-w+1}{s})$
$M \leftarrow zero(num+1, N, C)$
$i \leftarrow 1$ // window number
while ($i \le num$) do
    $W_i \leftarrow MSA(X_{(T2,N-(i-1)s-w+1)}^{l}; \ldots; X_{(T2,N-(i-1)s)}^{l})$
    $M[i-1, N-(i-1)s-w : N-(i-1)s-1, :] \leftarrow W_i$
    $i \leftarrow i + 1$
end while
if $(N - (num-1)s - w + 1 ! = 1)$ then // complete window are divided and redundant tokens exist
    $W_i \leftarrow MSA(X_{(T2,1)}^{l}; \ldots; X_{(T2,N-(i-1)s)}^{l})$
    $M[i-1, 0 : N - num * s - 1, :] \leftarrow W_i$
end if
$count \leftarrow (M[:, :, C] ! = zero(C)).sum(axis = 0)$ // count the number of overlapping tokens for each location
$t \leftarrow M.sum(axis = 0)/count$ // average to get the final output
$X_{T2}^{l+1} \leftarrow MLP(LN(t)) + t$
return $X_{T2}^{l+1}$
```

4. Results

4.1. Datasets

This study used Alzheimer's Disease Neuroimaging Initiative (ADNI) [36] database (http://adni.loni.usc.edu) to verify the performance of the proposed model. Founded in 2003, led by Principal Investigator Michael W. Weiner, MD, the ADNI is a publicly available clinical and imaging database for AD. ADNI is testing more than 5000 subjects from 59 locations around the world. Its primary goal is to diagnose MCI and early AD by taking magnetic

Table 1
Configuration details of four sliding-window attention blocks of token series T1 and token series T2.

Block num	Block name	Window sliding orientation	Window size	Sliding stride
2	RSwin	right	5	4
4	LSwin	left	9	8
6	RSwin	right	17	16
8	LSwin	left	33	32

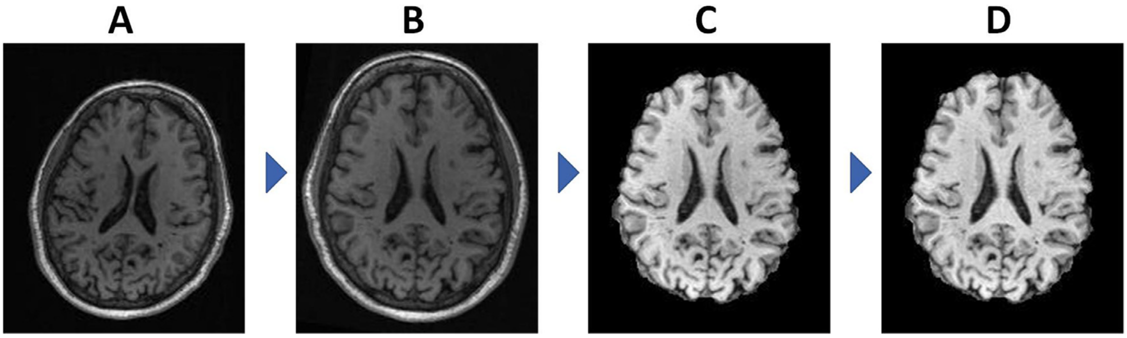

Fig. 6. Brain sMRI image preprocessing pipline (axial view). (A) Original brain sMRI image. (B) Space normalization. (C) Skull dissection. (D) N4 bias field correction.

Table 2
Subjects demographics.

Category	Number (subjects)	M/F (subjects)	Number (samples)	M/F (samples)	Age
sMCI	154	84/70	449	255/194	74.8±4.1
pMCI	121	63/58	374	190/184	75.2±5.6

resonance imaging (MRI), positron emission tomography (PET) and other biomarkers, combined with clinical and neuropsychological assessments.

Since the sMRI images adopted by this study are from ADNI 1/GO/2/3 stage, in order to avoid the influence of factors other than brain structure on the model, our research is based on the preprocessed MRI. Firstly, FMRIB Software Library (FSL) was used to normalize the sMRI images into MNI152 standard space. After normalization, all images had a dimension of 182 × 218 × 182 $(X \times Y \times Z)$ with a spatial resolution of $1 \times 1 \times 1mm^3$ per voxel, then skull dissection was performed on the spatial normalized sMRI images, finally unified bias field correction was performed using Advanced Normalization Tools (ANTs). Image preprocessing pipline is shown in Fig. 6. Pre-processed sMRI images were sliced according to the axial direction. When selecting the axial slices of each MRI image, we sliced from the middle to both ends in the vertical axial plane direction, that is, 40 axial slices were cut from the 91st aixal slice to each end, for a total of 80 slices. Finally, each sample contains two 3D sMRI images: sMRI image T1 and sMRI image T2, which correspond to slice series T1 and slice series T2 after slicing. Slice series T1 and slice series T2 have the same number of slices, and the spatial dimension of each slice is 182 ×218 × 1.

For a single sample, if the patient was diagnosed with MCI at T2 and did not progress to AD in the next 3 years, we divided the sample into stable MCI (sMCI) group. If the patients was diagnosed with MCI at T2 and conversion to AD occurred during the following 3-year follow-up, the sample was divided into the progressive MCI (pMCI) group. Since a patient may have sMRI images acquired at multiple time points, each patient may correspond to multiple samples. After condition screening, 823 samples were collected from 275 subjects. Subjects demographics are shown in Table 2, the T2 moment in each sample is taken as the collection time of this sample.

4.2. Experimental setup

The 823 samples were divided into three subsets: training, validation and test, of which 65% for training, 20% for validation and 15% for test. In addition, we must ensure that samples in different subsets cannot come from the same subject, in other words, multiple samples from the same subject are only allowed to exist in one of the training, validation and test subsets. This is very important because samples belonging to the same subject have a high similarity in image structure, which will lead to overfitting of the model. Additionally, mirror flipping and random rotation operations were used to augment the dataset. For the two slice series in each sample, due to the GPU limitations, we set the number of slices N = 80 in each slice series, and the token dimension C = 256. The model was trained with a learning rate of $1e^{-5}$ and momentum of $9e^{-5}$ for 100 epochs. To avoid overfitting, we chose SGD [37] as optimizer with a weight decay of 0.1. We used cross-entropy [38] as loss function. To avoid randomness, this study conducted 5 experiments, each experiment randomly selected different training, validation and test subset.

4.3. Performance evaluation and comparison

This paper uses accuracy, sensitivity, specificity, receiver operating curve (ROC), and area under ROC curve (AUC) as model performance evaluation criteria. Label true positive, true negative, false positive, and false negative as TP, TN, FP and FN, then accuracy, sensitivity, specificity can be expressed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

Table 3
The comparative analysis of the various indicators using the proposed model with the recent deep learning methods on pMCI vs sMCI task.

Methods	Techniques	Convert time (month)	Accuracy(%)	Sensitivity(%)	Specificity(%)	AUC
Suk et al. [14]	2D CNN	0–18	74.82	70.93	78.82	0.7539
Basaia et al. [15]	3D CNN	0–36	75.10	74.80	75.30	-
Cui et al. [20]	3D CNN+RNN	-	71.71	65.27	76.27	0.7303
Li et al. [16]	3D CNN+RNN	-	72.50	61.00	82.50	0.7460
Oh et al. [17]	3D CNN	0–36	73.95	77.46	70.71	0.7911
Base 3D model [18]	3D CNN	0–36	71.07	71.41	64.15	0.7538
3D Recurrent Attention model [18]	3D CNN+RNN	0–36	74.00	75.09	71.14	0.7797
Attention Transformer model [18]	2D CNN+Transformer	0–36	75.34	76.25	72.79	0.7926
VGG-TSwinformer	2D CNN+Transformer	0–36	77.20	79.97	71.59	0.8153

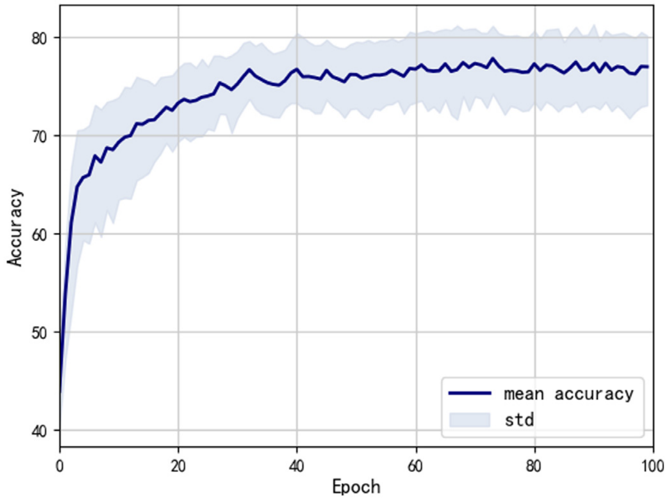

Fig. 7. Average accuracy curve of the validation set.

$$Specificity = \frac{TN}{FP + TN} \tag{14}$$

In addition to high accuracy, we seek higher sensitivity and specificity, but we usually need to find a balance point between the sensitivity and the specificity, and the ROC curve can represent this process. Furthermore, we use AUC to evaluate the performance of the model.

Fig. 7 shows the average accuracy and standard deviation for the validation dataset during training. It can be seen that the model validation accuracy has converged since the 60th epoch. In this study, the "Base 3D Model", "3D Recurrent Visual Attention Model" and "Attention Transformer Model" proposed in [18] are compared with the proposed VGG-TSwinformer. Since temporal attention is not adopted in [18], so when these three models are applied to the dataset for this study, only the slice series T2 in each sample is used. Five random experiments were conducted for each model, and the random seeds were consistent with those of VGG-TSwinformer, and the average of test accuracy, sensitivity, specificity and AUC of three contrasting models and VGG-TSwinformer are shown in Table 3, and mean ROC curves of three contrasting models and VGG-TSwinformer are given in Fig. 8. In addition, we directly compare the performance of recent sMRI-based studies [14–17,20] in the classification task of pMCI and sMCI in Table 3. In order to eliminate the algorithm performance bias due to different datasets, we only compare the algorithms applied on the ADNI dataset. From the comparison of experimental results in Table 3 we can see that the proposed model is superior to the deep learning algorithms based on cross-sectional research in the accuracy, sensitivity and AUC, achieving 77.2%, 79.97% and 0.8153, respectively. It is worth mentioning that the AUC and sensitivity of this method are significantly better than other methods, which are 2.51 and 2.27 percentage points higher than the second place, respectively.

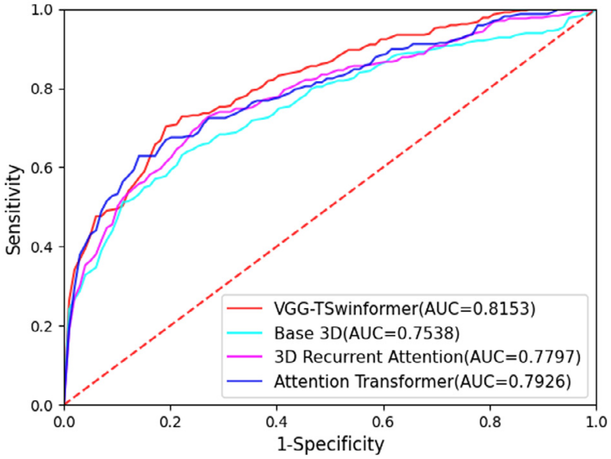

Fig. 8. Average ROC curve of baseline models and VGG-TSwinformer.

Fig. 9 shows the attention areas of the trained VGG-TSwinformer model for sMRI image slices in multiple samples, we produced it by applying Grad-CAM [39] for the last convolution layer after the second pooling layer in VGG-16. Analysis of the heatmaps reveals that the most commonly affected brain areas are around the temporal lobe, including the amygdala and hippocampus, suggesting that changes in the amygdala and hippocampus play an important role in the progression of MCI. The second commonly affected areas are the angular gyrus, precuneus, and middle temporal gyrus, etc., but they are only present in some patients and are less common than the hippocampus. In addition, in the attention heat maps of some correctly classified samples, the signs of the above common discriminative areas are not obvious, which indicates that some brain functional areas other than these common discriminative areas also have auxiliary discriminative functions for distinguishing pMCI and sMCI. It should be noted that the formation of the highlighted areas in the heat map is not only due to the longitudinal changes of brain structure during T1 and T2, but also related to the spatial characteristics of brain structure at T1 and T2. VGG-TSwinformer uses these two features to achieve competitive performance on the MCI progress prediction task, but does not distinguish them. It is of great significance to visually display the progression of MRI using longitudinal changes in brain morphology, which helps clinicians to grasp the changes in the local brain of MCI patients over a certain period of time. This can be explored as another direction.

4.4. Model performance analysis

To demonstrate that our model is more sensitive to the changing brain structures, we conducted controlled experiments. We

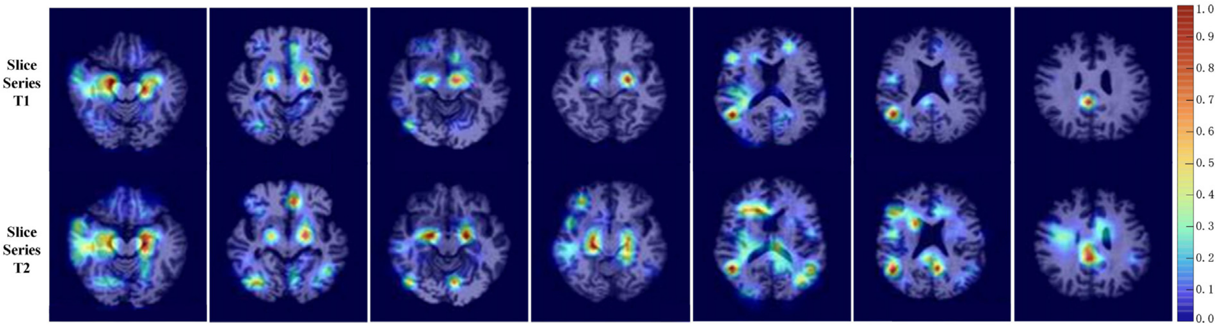

Fig. 9. Network attention areas for sMRI image slices.

Table 4
Comparison of controled experiment and original experiment.

	Original experiment	Controled experiment
Accuracy	0.7720	0.7276
Sensitivity	0.7997	0.7533
Specificity	0.7159	0.7346
AUC	0.8153	0.7843

Table 5
Comparison of pre-trained VGG-TSwinformer and our VGG-TSwinformer.

	VGG-TSwinformer trained from scratch	VGG-TSwinformer with pre-trained VGG-16
Accuracy	0.7720	0.7622
Sensitivity	0.7997	0.8132
Specificity	0.7159	0.6663
AUC	0.8153	0.7953

kept the model architecture unchanged and only replaced all slices in slice series T1 with slices in slice series T2 in each sample, that is, each sample contains two slice series that are exactly the same. We conducted 5 controlled experiments, and the dataset division of each controlled experiment was consistent with that of the original experiment. The average accuracy, sensitivity, specificity and AUC obtained are shown in Table 4.

It can be seen that the performance of the model is not competitive with other algorithms when the samples no longer contain the information of changes in the brain structure of MCI patients, which indicates that VGG-TSwinformer can utilize information on structural changes in the patient's brain to improve performance. In addition, the sensitivity of the original experiment is higher than that of the control experiment, indicating that the model is more sensitive to the changing brain structures. However, the specificity of the original experiment is lower than that of the control experiment. Higher sensitivity means lower rates of missed diagnosis, in fact, in the vast majority of clinical diagnosis, the missed diagnosis is more unacceptable than the misdiagnosis.

To investigate whether using a pre-trained CNN will improve the model performance, we used VGG-16 loaded with pre-trained weights in VGG-TSwinformer and compared it with the model trained from scratch. As shown in Table 5, the accuracy, specificity and AUC of the model trained from scratch outperform the pre-trained model, proving that using pre-trained VGG does not improve model prediction performance.

MRI has three plane views: axial (divides the body into top and bottom halves), coronal (perpendicular), and sagittal (midline of the body). We compared the performance of the model using different plane slices. For the coronal plane, 40 slices were taken from the 109th coronal slice to each end and a total of 80 slices were got in each sMRI image; for the sagittal plane, 40 slices were taken from the 91th sagittal slice to each end and a total of 80 slices were got. It can be seen from Table 6 that the effect of the model will vary according to the choice of plane. In general, the model has the best comprehensive performance when used axial plane slices, but the sensitivity and specificity are lower than those of coronal plane and sagittal plane respectively. Therefore, only using a single plane slice cannot fully extract the features of 3D MRI, and combining the three plane slices may achieve better performance.

Table 6
Comparison of VGG-TSwinformer on three plane slices of axial, coronal and sagittal.

	Axial plane	Coronal plane	Sagittal plane
Accuracy	0.7720	0.7548	0.7594
Sensitivity	0.7997	0.8079	0.7348
Specificity	0.7159	0.7026	0.7669
AUC	0.8153	0.7904	0.7778

5. Conclusion

This paper propose a novel deep learning model VGG-TSwinformer for longitudinal study of MCI. In this model, VGG-16 based CNN is used to extract features from sMRI image slices and encodes them as high-level feature representation tokens. Temporal attention mechanism is used to link longitudinal sMRI images to obtain information on brain structural changes in MCI patients. The sliding-window attention mechanism can fully integrate the local spatial features of sMRI image, and gradually integrate the remote spatial features through the superposition of attention windows of different sizes. We validated our model performance on the ADNI database, and the experimental results shown that the proposed VGG-TSwinformer achieved better diagnostic efficiency compared to sMRI-based cross-sectional studies.

For the progress prediction task of MCI, using longitudinal biomarkers can avoid the shortcomings of cross-sectional studies that are affected by individual differences and short observation time, and lead to better diagnostic efficiency. However, there are few applications of deep learning for longitudinal analysis of MCI due to the inherent disadvantage of deep learning in dealing with longitudinal medical image data. Encouragingly, Transformer's success in computer vision may advance the application of deep learning in short-term longitudinal disease studies. Despite the competitive performance on the MCI prediction task, this study still suffers from the following limitations. First, this study uses MRI slices as local information of sMRI image, and does not mine 2D local features inside slices. Second, this work investigates model performance on axial, coronal and sagittal plane slices, but does not adopt an effective feature fusion method to fuse these three plane features, which will cause the loss of sMRI spatial information. Third, since the proposed VGG-TSwinformer only use sMRI as the only biomarker and does not take full advantage of the available cross-sectional biomarkers, resulting in the VGG-TSwinformer model may not achieve the corresponding diagnostic efficiency compared to those using multimodal biomarkers, and combining multimodal cross-sectional and longitudinal biomarkers is also the goal of our future research.

Declaration of Competing Interest

The authors declared no conflict of interest.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 61976080), the Science and Technology Key Project of Science and Technology Department of Henan Province (Grant No. 212102310298), the Academic Degrees & Graduate Education Reform Project of Henan Province (Grant No. 2021SJGLX195Y), the Innovation and Quality Improvement Project for Graduate Education of Henan University (SYL20010101, SYLYC2022191, SYLYC2022192).

Data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The investigators within the ADNI did not participate in analysis or writing of this study. A complete list of ADNI investigators can be found online.

References

[1] J. Gaugler, B. James, T. Johnson, J. Reimer, M. Solis, J. Weuve, R.F. Buckley, T.J. Hohman, 2022 Alzheimer's disease facts and figures, Alzheimer's Dement. 18 (2022) 700–789.

[2] B. Khagi, G.R. Kwon, 3D CNN design for the classification of Alzheimer's disease using brain MRI and PET, IEEE Access. 8 (2020) 217830–217847.

[3] I. Mebane-Sims, Alzheimer's association, 2018 Alzheimer's disease facts and figures, Alzheimers Dement 14 (3) (2018) 367–429.

[4] B.J. Kelley, R.C. Petersen, Alzheimer'S disease and mild cognitive impairment, Neurol. Clin. 25 (2007) 577–609.

[5] C. Sarmiento, C. Lau, Diagnostic and statistical manual of mental disorders: DSM-5, Wiley Encycl. Personal. Individ. Differ. Personal. Process. Individ. Differ. (2020) 125–129.

[6] C.R. Jack, M.M. Shiung, S.D. Weigand, P.C. O'brien, J.L. Gunter, B.F. Boeve, D.S. Knopman, G.E. Smith, R.J. Ivnik, E.G. Tangalos, Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI, Neurology. 65 (2005) 1227–1231.

[7] L.-M. Bogza, C. Patry-Lebeau, E. Farmanova, H.O. Witteman, J. Elliott, P. Stolee, C. Hudon, A.M.C. Giguere, User-centered design and evaluation of a web-based decision aid for older adults living with mild cognitive impairment and their health care providers: mixed methods study, J. Med. Internet Res. 22 (2020) e17406.

[8] J. Samper-Gonzâlez, N. Burgos, S. Bottani, S. Fontanella, P. Lu, A. Marcoux, A. Routier, J. Guillon, M. Bacci, J. Wen, Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data, Neuroimage. 183 (2018) 504–521.

[9] M.A. Ebrahimighahnavieh, S. Luo, R. Chiong, Deep learning to detect alzheimer's disease from neuroimaging: a systematic literature review, Comput. Methods Programs Biomed. 187 (2020) 105242.

[10] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, X. Zhang, Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease, Neurocomputing. 361 (2019) 185–195.

[11] C. Yin, S. Li, W. Zhao, J. Feng, Brain imaging of mild cognitive impairment and Alzheimer's disease, Neural Regen. Res. 8 (2013) 435.

[12] L.I. Kun-cheng, Progress of neuroimaging research on Alzheimer's disease, Chinese J. Contemp. Neurol. Neurosurg. 14 (2014) 176.

[13] G. Martí-Juan, G. Sanroma-Guell, G. Piella, A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease, Comput. Methods Programs Biomed. 189 (2020) 105348.

[14] H.-I. Suk, S.-W. Lee, D. Shen, The Alzheimer's disease neuroimaging initiative (ADNI), deep ensemble learning of sparse regression models for brain disease diagnosis, Med. Image Anal. 37 (2017) 101–113.

[15] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, The Alzheimer's disease neuroimaging initiative (ADNI), automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, NeuroImage Clin. 21 (2019) 101645.

[16] F. Li, M. Liu, The Alzheimer's disease neuroimaging initiative (ADNI), a hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease, J. Neurosci. Methods. 323 (2019) 108–118.

[17] K. Oh, Y.-C. Chung, K.W. Kim, W.-S. Kim, I.S. Oh, Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning, Sci. Rep. 9 (2019) 1–16.

[18] F. Altay, G.R. Sânchez, Y. James, S.V. Faraone, S. Velipasalar, A. Salekin, Preclinical stage Alzheimer's disease detection using magnetic resonance image scans, in: Proc. AAAI Conf. Artif. Intell., 2021, pp. 15088–15097.

[19] M. Aghili, S. Tabarestani, M. Adjouadi, E. Adeli, Predictive modeling of longitudinal data for Alzheimer's disease diagnosis using RNNs, in: Int. Work. Predict. Intell. Med., 2018, pp. 112–119.

[20] R. Cui, M. Liu, The alzheimer's disease neuroimaging initiative (ADNI), RNN-based longitudinal analysis for diagnosis of alzheimer's disease, Comput. Med. Imaging Graph. 73 (2019) 1–10.

[21] R. Cui, M. Liu, G. Li, Longitudinal analysis for Alzheimer's disease diagnosis using RNN, in: IEEE 15th Int. Symp. Biomed. Imaging (ISBI), 2018, pp. 1398–1401.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 213–229.

[24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H.S. Torr, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 6881–6890.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: transformers for image recognition at scale, ArXiv Prepr. ArXiv2010.11929 (2020).

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, Int. Conf. Mach. Learn. (ICML) (2021) 10347–10357.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.

[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90.

[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Int. Conf. Learn. Rep. (ICLR), 2022.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[32] Y. LeCun, P. Haffner, L. Bottou, Y. Bengio, Object recognition with gradient-based learning, in: Shape, Contour Group. Comput. Vis., 1999, pp. 319–345.

[33] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: transformers make strong encoders for medical image segmentation, ArXiv Prepr. ArXiv2102.04306 (2021).

[34] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, T. Liu, On layer normalization in the transformer architecture, Int. Conf. Mach. Learn. (ICML) (2020) 10524–10533.

[35] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), ArXiv Prepr. ArXiv1606.08415 (2016).

[36] C.R. Jack Jr, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, The alzheimer's disease neuroimaging initiative (ADNI): MRI methods, J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn. Reson. Med. 27 (2008) 685–691.

[37] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. (1951) 400–407.

[38] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3D brain MRI classification, in: IEEE 14th Int. Symp. Biomed. Imaging (ISBI), 2017, pp. 835–838.

[39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.