An Efficient Lightweight Network Based on Magnetic Resonance Images for Predicting Alzheimer’s Disease

Boan Ji, Anhui University, China
Huabin Wang, Anhui University, China*
Mengxin Zhang, Anhui University, China
Borun Mao, Anhui University, China
Xuejun Li, Anhui University, China

ABSTRACT

Brain magnetic resonance images (MRI) are widely used for the classification of Alzheimer’s disease (AD). The size of 3D images is, however, too large. Some of the sliced image features are lost, which results in conflicting network size and classification performance. This article uses key components in the transformer model to propose a new lightweight method, ensuring the lightness of the network and achieving highly accurate classification. First, the transformer model is imitated by using image patch input to enhance feature perception. Second, the Gaussian error linear unit (GELU), commonly used in transformer models, is used to enhance the generalization ability of the network. Finally, the network uses MRI slices as learning data. The depthwise separable convolution makes the network more lightweight. Experiments are carried out on the ADNI public database. The accuracy rate of AD vs. normal control (NC) experiments reaches 98.54%. The amount of network parameters is 1.3% of existing similar networks.

KEYWORDS
Alzheimer’s Disease Prediction, Computer-Aided Diagnosis, Convolutional Neural Network, Depthwise Separable Convolution, Magnetic Resonance Images

1. INTRODUCTION

Alzheimer’s disease (AD) is an incurable central neurodegenerative disorder that tends to occur in old age (Mayeux, 2010). The number of people with AD has increased dramatically due to the aging population. It is expected that there will be 60 million AD patients worldwide in 50 years (Alzheimer’s Association, 2019). AD is incurable; therefore, the best strategy to control AD is to identify people at higher risk of developing the disease as early as possible and intervene to prevent its effects. Due to the unclear etiology of AD, clinical diagnostic methods are limited to neuropsychological tests (Kowoll et al., 2015; Zhang et al., 2019).

Mild cognitive impairment (MCI) is a precursor to dementia. Patients with MCI present with moderate symptoms like mild memory loss. The symptoms do not affect independent living, which makes it more difficult to detect and, in turn, more likely to develop into AD. A growing number of
studies have shown that magnetic resonance imaging (MRI) can observe progressive brain loss in patients with MCI to complete AD (Whitwell et al., 2008).

Higher imaging parameters and high soft tissue resolution of the MRI compares to computed tomography (CT) images. MRI is the most used diagnostic image of the brain due to its accurate display of the 3D structure of the brain (see Figure 1). MRIs are widely used in clinical applications as an adjunct to neuropsychological testing for organic lesions. However, in clinical applications, the judgment of patients’ MRI is highly dependent on the physician’s consulting experience. This leads to the judgment of the patients’ condition being limited to the ability of the physician.

The convolutional layer is the most critical component of Convolution Neural Networks (CNNs). As a feature extractor, the convolutional kernel works by sweeping through the input features in a regular manner, summing the matrix elements and superimposing the bias amount within the perceptual field. This architecture introduces the inductive bias of locality and spatial invariance to CNNs, allowing CNNs to excel in image processing tasks. The sliding feature extraction approach uses convolutional kernels smaller than the feature map size, which allows the convolutional kernels to extract local features at one time. In addition, key feature changes to any position within the feature map can be sensed during the sliding process. Such an operation allows neurons to only connect to a small region of the previous layer of the feature map. This greatly reduces the number of parameters in the network (Albawi et al., 2017).

Several approaches have been used to classify AD via convolutional neural networks. Zhang et al. (2019) used a classification network with a deep small convolutional kernel to classify AD via sliced images. Li et al. (2017) proposed a multimodal network with a combination of 2D and 3D convolution for the classification of Alzheimer’s patients and normal individuals. However, the convolution operation focuses on local feature extraction, and the MRI global association information in the image is easily lost. Therefore, this article attempts to implement the classification of AD using the transformer model.

The transformer, a deep learning architecture initially applied to natural language processing (NLP), centers on extracting features using the self-attention mechanism. However, the transformer model is also applicable in vision tasks in which the self-attention mechanism adds a global perceptual domain to the model. It can learn image features autonomously without bias, enabling it to achieve performance beyond that of CNNs in several vision tasks (Han et al., 2021). Dosovitskiy et al. (2020) proposed a ViT model that splits the image into several patches, inputs the token obtained by path embedding into a pure transformer encoder, and uses the encoder post result for classification. This achieves a better classification performance than CNNs. In addition, the Swin transformer model proposed by Liu et al. (2021) added a movable window design to the ViT model. This restricted the attention mechanism to a nonoverlapping local window, beating all models at that time in the task of target detection and segmentation of multiple datasets.
The authors constructed several visual transformer models for AD classification experiments to capture feature information at a greater distance on sliced images that have lost spatial information with the global attention of the transformer model. The experimental results were inferior to those of existing convolutional neural network-based classification networks because transformer models have an obvious drawback. The self-attention mechanism does not have the inductive bias of localization and spatial invariance in CNNs, resulting in transformer models requiring extremely large amounts of data for training or using pre-trained weights when applied to CV. In other words, ViT must be pre-trained on 300 million images to achieve accuracy beyond that of ResNets (Dosovitskiy et al., 2020; He et al., 2016).

This drawback is fatal in AD prediction classification tasks with small datasets. To introduce the inductive bias of convolutional networks into the network, this article sets up experiments to “partially” introduce the inductive bias of CNNs into the transformer by using the GPSA module in ConViT (d’Ascoli et al., 2021). Still, the results are unsatisfactory in AD classification experiments. This indicates that inductive bias is important in classification tasks with small training sets. Therefore, the authors continued to use the CNNs framework for the predictive classification of AD. In addition, they introduced the processing method used in the transformer model to achieve an innovative approach to CNNs. This achieved satisfactory results in its experiments.

First, this article proposes to migrate the operation of segmenting the input image into patches in the vision transformer and cooperate with the large-size convolutional kernel to improve the feature extraction capability of the network. Second, this article proposes to replace the rectified linear unit (ReLU) in traditional convolutional networks with the Gaussian error linear unit (GELU). This is often used as the activation function in visual transformers to improve the generalization ability of the network. Third, this article proposes an AD classification network based on MRI slices. It uses deep separable convolution rather than standard convolution so the network can be lightweight while ensuring high classification accuracy.

2. RELATED WORK

Many teams have explored the use of MRI for the classification of AD. Feng et al. (2020) proposed a classification network using 3D MRI as a basis for classification. The classification network, which used 3D convolution with support vector machines, had a strong performance in binary classification experiments for AD (99.10% accuracy in AD vs. NC tasks and almost 99% sensitivity and specificity). Yu and Liao (2020) proposed an AD classification algorithm for stereo images using 3D-ResNet. Their work achieved 97.425% accuracy in an Alzheimer’s triple classification experiment with more than 85 million parameters. Their comparison experiment with 2D sliced images had an accuracy of 49.356% with the classification network using 2D-ResNet. This included half the number of parameters of 3D-ResNet. Similarly, Gunawardena et al. (2017) used a 2D slice-extracted segmented brain MRI with a 2D convolutional network to achieve maximum accuracy of 96% in their experiments. In the studies of Feng et al. (2020) and Yu and Liao (2020), AD diagnosis network tests were applied to 3D and 2D nuclear magnetic images respectively. The comparison of experimental results is shown in Table 1.

As can be seen from Table 1, studies reflect a conflict between network classification accuracy and network lightweight. This is shown by the fact that the accuracy and number of parameters of the 3D MRI-based Alzheimer’s classification network are significantly higher than those of the 2D MRI-based network. The use of sliced images can effectively reduce the number of network parameters, however, there will be some spatial feature loss that can lower the performance of the 2D classification network and 3D classification network.

This article, therefore, proposes a brain MRI slice-based AD classification network with high classification accuracy. It transfers the key components in the transformer model to improve the feature extraction ability of the model. In addition, it uses depthwise separable convolution instead of standard convolution to achieve lightweight.
Applying the transformer to vision necessitates controlling its computational cost. Applying self-attention to each pixel generates more computation than the available computational power. Therefore, in ViT, the authors partitioned the images into patches and applied attention directly to each patch. Several subsequent works, such as MLP-Mixers (Tolstikhin et al., 2021) and ConvMixer (Trockman & Kolter, 2022) demonstrated that the operation of segmenting input images would give a large performance gain to the model due to larger blocks of image patches effectively increasing the perceptual field of the network compared to the size of the convolutional kernel. It, in turn, allows it to mix more distant spatial information. In the experiments of Tolstikhin et al. (2021), the authors achieved excellent results in the image classification task using only the MLP network paired with the input image segmentation operation. This indicates that the segmentation operation effectively improves the classification performance.

This article also uses a larger convolutional kernel than other popular convolutional networks. Large convolutional kernels were once commonly used in convolutional networks. However, since VGGNet proposed replacing large convolutional kernels with multiple 3x3 kernels to improve the depth of the network (Simonyan & Zisserman, 2014), CNNs have almost all been designed with small convolutional kernels to improve the depth of the model. In the visual transformer, however, the model benefits from the global perception capability of the self-attention mechanism. It has better spatial fusion and feature extraction capability. Even the Swin transformer, which adopts a local window mechanism, still uses a large 7x7 window (Yu & Liao, 2020). This allows cross-window connections to equalize this spatial perception capability at the same time, which requires dressing a larger convolutional kernel.

In addition, in transformer networks like BERT (Devlin et al., 2018) and ViT (Dosovitskiy et al., 2020), the activation function is no longer chosen from ReLU, which is common in CNNs. However, it uses GULEs. GELUs, an improvement on ReLU, are a combination of dropout operation and ReLU. ReLU will deterministically multiply the input by 0 or 1. Dropout will randomly multiply by 0. GELU chooses to multiply by 0 or 1, depending on how probable the current input is greater than the other inputs. Therefore, its nonlinear variation is a stochastic canonical transformation in the expected way.

3. METHODS OF ALZHEIMER’S PREDICTION

3.1 Datasets

The data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership under the leadership of Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biomarkers, and clinical and neuropsychological assessments can be combined to measure mild cognitive impairment (MCI) progression and early stages of Alzheimer’s disease (AD). See www.adni-info.org for the latest information.

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of Parameters (M)</th>
<th>Accuracy (%)</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yu and Liao (2020)</td>
<td>85</td>
<td>97.43</td>
<td>3D-ResNet-101</td>
</tr>
<tr>
<td></td>
<td>42.50</td>
<td>49.36</td>
<td>2D-ResNet-101</td>
</tr>
<tr>
<td>Feng et al. (2020)</td>
<td>-</td>
<td>92.11</td>
<td>3D-CNN-SVM</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>82.57</td>
<td>2D-CNN</td>
</tr>
</tbody>
</table>
The experiments in this article use T1-weighted MP-RAGE sequence MRI from 180 subjects in ADNI as training and testing data. To avoid the influence of volunteers’ gender and age on the classification results, the experiments screen the same number of subjects per category with a similar average age (see Table 2).

Table 2. Subjects’ pathological information statistics

<table>
<thead>
<tr>
<th>Label</th>
<th>AD</th>
<th>MCI</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Gender(M/F)</td>
<td>40/40</td>
<td>40/40</td>
<td>40/40</td>
</tr>
<tr>
<td>Age</td>
<td>76.48±8.34</td>
<td>76.08±7.96</td>
<td>76.66±7.07</td>
</tr>
<tr>
<td>MMSE</td>
<td>23.44±1.9</td>
<td>27.68±2.0</td>
<td>28.71±1.7</td>
</tr>
</tbody>
</table>

In the table, MMSE indicates the subject’s simple mental state examination scale score. It is a comprehensive and accurate reflection of the subject’s mental state and degree of cognitive deficits. The test scores are influenced by education level.

The MP-RAGE sequence MRI in this article was gradient distorted, intensity corrected, and gradient drift scaled using phantom data. In addition, these images were normalized and mapped onto a standard TPM template using SPM12 software, which is commonly used for biological neuroimaging processing.

3.2 Efficient Lightweight Classification Network

The AD prediction network proposed in this article can be divided into three parts. The first part is the patch embedding of the input MRI. The network uses disjoint Patch-sized convolution kernels to divide the input image into patches. It embeds the patches into the network. The second part is composed of residual connected convolution blocks. The convolution kernels are equipped with depth-separable convolution. It uses Depthwise convolution and Pointwise convolution instead of standard convolution. It sets GELU activation after each layer of Deepwise convolution and batch normalization after each layer of Pointwise convolution. This design is a transformer style with less activation and normalization compared to ResNet’s convolutional block. In addition, at the end of the second part, we set up an adaptive average pooling layer to pool the 16×16 feature maps obtained from the convolution to obtain smaller feature maps. The pooling layer design will significantly reduce the number of parameters in the third fully connected layer to meet the requirements of the lightweight design of the network. The pooling layer design can extract higher dimensional feature information from the feature maps. The third part of the network is the fully connected layer. It flattens the pooled feature maps and feeds them into the fully connected network to complete the classification. Figure 2 illustrates the various parts of the network.
First, before feature extraction, the input image must be sliced and diced into patches of larger sizes. The implementation method is like the chunking method in ViT, in which the input image is chunked and encoded using a 2D convolution operation with a convolution kernel size equal to the step size. Figure 3 shows how to chunk the input image.

In addition, this article uses GELUs to replace the operation of conventional convolutional neural networks via ReLU functions with a dropout layer (Hendrycks & Gimpel, 2016). The form of the GELU function is shown here:

\[
GELU(x) = xP(X \leq x) = x\phi(x)
\]

where \( P(X \leq x) \) refers to the probability that \( X \) is less than or equal to a given value of \( x \), and \( \phi(x) \) refers to the cumulative distribution of the Gaussian normal distribution of \( x \). In other words:

\[
\phi(x) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt
\]

That is, GELU can be reduced by the tanh approximation to:
\[ GELU(x) = xP\left( X \leq x \right) = 0.5 \times x \left[ 1 + \tanh \left( \frac{2}{\sqrt{\pi}} \left( x + 0.044715x^3 \right) \right) \right] \]

Therefore, the probability \( P\left( X \leq x \right) \) (i.e., the Gaussian normal distribution \( \phi(x) \) of \( x \)) varies with \( x \). When \( x \) is smaller, the more likely the activation result will be 0 in the case of the current activation function. The neuron is dropout. When \( x \) is larger, the more likely it is to be retained (Hendrycks & Gimpel, 2016).

The design of the network convolution module sets fewer normalization and activation operations. Compared to the design in ResNets, this article halves both operations (He et al., 2016). It is also inspired by the Swin transformer, which uses one activation and one normalization operation per convolutional module.

Finally, the authors choose to use depthwise separable convolution instead of the standard convolution approach (Chollet, 2017). The depthwise convolution and pointwise convolution fuse the feature maps spatially and channelwise, respectively. Such an operation will facilitate the lightweight design of the network. It is used by lightweight CNNs like MobileNets (Howard, 2017).
The idea of deeply separable convolution, as in Figure 5, is to extract features using two independent steps of lightweight convolution. First, it extracts features on each channel separately. Then, it uses a size 1×1 convolution kernel in the second convolution layer to achieve feature fusion across channels. In addition, this design ensures that the network is lightweight without losing the ability to perceive the features.

Assuming that the input feature map size is \( D_{in} \times D_{in} \times C_{in} \) (\( D_{in} \) is the input feature map size, \( C_{in} \) is the number of input channels), then there are \( C_{in} \times K \times K \) convolution kernels for depthwise convolution output size. There are \( C_{out} \times 1 \times 1 \) convolution kernels for pointwise convolution. The total number of parameters can be calculated from Equation (4):

\[
\text{The total number of parameters} = K \times K \times C_{in} + K \times K \times C_{in} \times C_{out}
\]  

(4)

The number of parameters for standard convolution using the \( K \times K \times C_{out} \) convolution kernel can be calculated from Equation (5):

\[
\text{The total number of parameters} = K \times K \times C_{in} \times C_{out}
\]  

(5)

In Equation (4) and Equation (5), \( D \) represents the size of the feature map. \( C \) represents the number of channels of the feature map. \( K \) represents the size of the convolution kernel.

Taking the parameters of the convolution kernel in this article as an example, the number of standard convolution kernel parameters is \( 7 \times 7 \times 512 \times 512 = 12845056 \). When using the separation operation, the number of parameters is \( 7 \times 7 \times 512 + 1 \times 1 \times 512 \times 512 = 287232 \). The number of standard convolution parameters is more than 400 times the depthwise separable convolution. By using this operation and combining it with the reduction of the feature map by dividing the input images, the network in this article achieves a lightweight.

4. EXPERIMENTS AND DISCUSSION

4.1 Experiments

In this section, a series of experiments are set up to test the AD diagnostic network proposed in this article. The experiments used the MRIs of 240 subjects from the ADNI dataset, 80 with AD, MCI, and normal control (NC). The MRI images were sliced horizontally into 2D images, as shown in Figure 6. A total of 4,320 2D MRI cross-sectional slices were obtained. 4,000 images with better slicing were manually screened. The remaining images were discarded. The size of the remaining images was corrected to 224×224. It was used as the final experimental data. The experimental data set was randomly divided into training and test sets. The performance of the proposed algorithm was tested in AD vs. NC, AD vs. MCI, MCI vs. NC, and AD vs. MCI vs. NC experiments, respectively. The experiments in this paper focus on the classification performance of the network and the number of parameters of the network. The amount of parameters can be calculated directly, so the authors set up a large number of experiments to verify the classification performance of the network. The authors use accuracy, sensitivity, precision, and specificity as metrics to measure the classification performance of the network.

The experiments in this article use the PyTorch deep learning framework. The experimental environment is Pytotch 1.10.2 and Torchvision 0.11.3. The computing hardware is NVIDIA RTX2080 GPU with CUDA version 11.4, Intel i7-8750H CPU, and 24 GB RAM.
The experiments set the model depth to 5 (N = 5) and the patch size to 14. The input image is segmented and encoded using a convolutional layer with a step size and convolutional kernel size of 14. The encoding dimension is set to 512. The feature map obtained by patch embedding is shown in Figure 7. The image has been segmented into patches; therefore, the feature map size does not need to perform downsampling processing (when patch size = 14, the feature map size is 16×16).

The cross-entropy function is chosen for the loss function during model training. The formula is shown here:

$$\text{Loss} = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^{M} y_{ic} \log(p_{ic})$$

(6)

$M$ in the formula denotes the number of categories for classification. $y_{ic}$ is the symbolic function (1 or 0), which takes 1 if $i$ and $c$ belong to the same group. Otherwise, $y_{ic}$ takes 0. $p_{ic}$ is the probability of predicting the $i$th sample as label $c$.

The learning rate decay during the training process takes the cosine annealing strategy. This strategy will effectively avoid training into local optimal solutions. The learning rate decreases in $T_{max}$ cycles according to a regular pattern. When the hot restart is performed after $T_{max}$ epochs, the learning rate will increase to the initial value. The increased initial value will help the network avoid becoming trapped in a local optimum solution:
\[ \eta_t = \frac{1}{2} (\eta_{\text{max}} - \eta_{\text{min}}) \left(1 + \cos \left(\frac{T_{\text{cur}}}{T_{\text{max}}} \pi\right)\right) \] (7)

For Equation (7), \( T \) is the calculated current learning rate. \( \eta_{\text{max}} \) and \( \eta_{\text{min}} \) represent the maximum and minimum values of the learning rate, respectively. They define the range of the learning rate. \( T_{\text{cur}} \) indicates how many epochs are currently executed. \( T_{\text{cur}} \) will be updated after each batch runs. \( T_{\text{max}} \) is the maximum number of epochs designed.

The following Figure 8 shows the loss and accuracy of the training process with AD vs. NC classification experiments, respectively.

Figure 8. Training loss and validation accuracy of AD vs. NC (fluctuations in black boxes are normal fluctuations due to hot restarts of the learning rate decaying according to the cosine annealing strategy)

Accuracy calculation formula:

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \] (8)

In this article, three binary classification and one ternary classification experiments were set up to verify the classification ability of the model. In addition to the accuracy of classification, the experiments need to focus on the sensitivity, precision, and specificity of classification. They are calculated as follows:

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \] (9)

\[ \text{Precision} = \frac{TP}{TP + FP} \] (10)

\[ \text{Specificity} = \frac{TN}{FP + TN} \] (11)
The abbreviations in the prior equation represents the correspondence between the true category of the test sample and the predicted value of the network output. The correspondence is shown in Table 3.

**Table 3. Meaning of the abbreviations in the Equations (9), (10), and (11)**

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Positive</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

When specifically combined with the diagnostic task of AD, the AD vs. NC classification accuracy, for example, indicates the probability of being able to correctly classify all AD patients and normal individuals separately. The sensitivity indicates the probability of being able to correctly diagnose AD. Accuracy indicates the probability that the subjects diagnosed with AD are indeed AD patients. Specificity indicates the probability of being able to correctly separate normal subjects. These four metrics are related to the four indicators that are positively correlated with the classification performance of the model. However, it is often difficult to ensure that both accuracy and sensitivity are relatively high. Therefore, the authors induced the F1-score as a reconciled average indicator of the two, with the following equation:

\[
F_1 \text{- score} = \frac{2 \times \text{Sensitivity} \times \text{precision}}{\text{Sensitivity} + \text{Precision}}
\]

**Table 4. Results of the four experiments**

<table>
<thead>
<tr>
<th>Task</th>
<th>Accuracy (%)</th>
<th>Specificity (%)</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD vs. NC</td>
<td>98.54</td>
<td>97.92</td>
<td>98.34</td>
</tr>
<tr>
<td>AD vs. MCI</td>
<td>91.35</td>
<td>89.17</td>
<td>91.07</td>
</tr>
<tr>
<td>MCI vs. NC</td>
<td>92.60</td>
<td>91.04</td>
<td>92.73</td>
</tr>
<tr>
<td>AD vs. MCI vs. NC</td>
<td>93.33</td>
<td>93.33</td>
<td>93.63</td>
</tr>
</tbody>
</table>

From the comparison of the data in Table 4, the strongest classification performance of the model in the dichotomous classification experiment was for AD patients and normal subjects, with an average accuracy of 98.54%. There were slightly lower results for AD/MCI and MCI/NC classification. Such experimental results are consistent with differences in brain structures of subjects with different conditions. MCI is an intermediate state between normal and Alzheimer’s patients, whose brain characteristics are different from those of the other two groups (differences are smaller and more difficult to classify). In the triple classification experiment, the analysis of the misclassified samples revealed that most of the samples belonged to AD/MCI and MCI/NC classification errors.
4.2 Discussion

The horizontal coordinate of the ROC curve is the false positive rate (FPR). The value is equal to 1-specificity. The vertical coordinate is the true positive rate (TPR), whose value is equal to accuracy. The closer the curve is to the upper left, the higher the performance of the classifier. The AUC value is the area enclosed by the curve and the coordinate axis. The size of the area directly reflects its classification ability.

To further verify the classification performance of the network proposed in this article, the authors draw the ROC curve of the classification experiment (see Figure 9).

Figure 9. ROC curve of the experiments

Computer-aided diagnosis of AD has received attention in both the computer and biomedical fields. Several teams have conducted recent research; however, each had different strategies for diagnosis. In this article, the research results of several teams within the last two years were selected to compare the performance of AD/NC classification with the method proposed in this article. The number of parameters of their proposed network was calculated to verify the lightness of the network in this article. The comparison results are shown in Table 5.

Table 5. Comparison of different methods

<table>
<thead>
<tr>
<th>Study</th>
<th>Number of Parameters (M)</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duc et al. (2020)</td>
<td>780.23</td>
<td>85.27</td>
<td>67.23</td>
<td>98.21</td>
<td>3D fMRI</td>
</tr>
<tr>
<td>Zhang et al. (2021)</td>
<td>64.13</td>
<td>91.3</td>
<td>91.0</td>
<td>91.9</td>
<td>3D MRI</td>
</tr>
<tr>
<td>Mehmood et al. (2021)</td>
<td>132.70</td>
<td>98.73</td>
<td>98.19</td>
<td>99.09</td>
<td>2D MRI</td>
</tr>
<tr>
<td>Odusami et al. (2021)</td>
<td>10.71</td>
<td>80.80</td>
<td>-</td>
<td>92.0</td>
<td>2D fMRI</td>
</tr>
<tr>
<td>Murugan et al. (2021)</td>
<td>4.53</td>
<td>95.23</td>
<td>-</td>
<td>95.0</td>
<td>2D MRI</td>
</tr>
<tr>
<td>Pradhan et al. (2021)</td>
<td>143.65</td>
<td>82.6</td>
<td>-</td>
<td>-</td>
<td>2D MRI</td>
</tr>
<tr>
<td>Kong et al. (2022)</td>
<td>-</td>
<td>93.21</td>
<td>91.43</td>
<td>95.42</td>
<td>3D MRI + 3D PET</td>
</tr>
<tr>
<td>Ouchicha et al. (2022)</td>
<td>73.26</td>
<td>98.53</td>
<td>-</td>
<td>-</td>
<td>2D MRI</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>1.74</td>
<td>98.54</td>
<td>98.78</td>
<td>97.92</td>
<td>2D MRI</td>
</tr>
</tbody>
</table>
In comparison with the existing work, the classification network of AD proposed in this article exceeds most of the existing studies in classification accuracy. The network is significantly ahead of the existing studies in lightness. In the table’s comparison, only the research of Mehmood et al. (2021) is higher in classification accuracy and specificity than the network proposed in this article, leading the accuracy by 0.2%. However, the parameter quantity of the network is 76 times that of this article. It is reasonable to think that the performance of the network is lower than that of the network proposed in this article when considering the lightweight of the network.

Similarly, the classification accuracy of the research of Ouchicha et al. (2022) in the table is like this article. However, the network parameter quantity is 42 times that of this article. In addition, there is a gap between the network and the network proposed in this article in terms of lightness. In conclusion, considering the accuracy of the classification and the lightweight of the network, the performance of the efficient AD classification network proposed in this article is optimal.

To verify the interpretability of the diagnostic network proposed in this article, the attentional heat map of the network during the test is visualized (see Figure 10). The red area is the network attentional concentration area. Three subjects with different stages are included in Figure 10: (a) AD subjects; (b) MCI subjects; and (c) NC subjects. In the diagnosis of clinical AD, neuroimaging prefers to make the diagnosis based on lesions in the bilateral temporal lobe, hippocampus, and posterior temporal lobe in the patient’s MRI. As seen in Figure 10(a), network attention is concentrated in the posterior temporal lobe region. There is atrophy of brain tissue visible to the naked eye in the most sensitive locations of the network. As seen in Figure 10(b), network attention is concentrated bilaterally in the temporal lobe regions. It is more concentrated in a relatively posterior position. In Figure 10(c), network attention is concentrated in the posterior temporal lobe area. It can be concluded that the sensitive regions of the diagnostic network proposed in this article are consistent with the regions focused in clinical diagnosis for the three stages of subjects. The predictive network proposed in this article is interpretable.

In addition, some teams have been exploring the use of positron emission tomography (PET) images in the field of AD-assisted diagnosis. PET, a newer type of medical imaging modality, injects radionuclide-labeled glucose and other substances necessary for biometabolism into the body. The aggregation of these substances in the metabolism reflects the metabolic activity of life. It, thus, completes the diagnosis.
Theoretically, the metabolic activity of the brain is altered in AD patients. PET imaging can also be used to aid in the diagnosis of AD. However, since PET imaging is an emerging technology, it is not widely available. Only high-level, large-scale hospitals have the relevant equipment. There is a significant cost to using the technology. This article only selects MRI as the experimental sample. It does not compare the performance with studies using PET images as the basis for classification.

To verify the efficiency of the proposed network in this article, the authors compared the running time of this method with commonly used image classification backbone networks like ResNet50, EfficientNetB5, VGG16, and ResNeXt50. The results of the comparison are shown in Table 6.

<table>
<thead>
<tr>
<th>Networks</th>
<th>Training time (m)</th>
<th>Testing time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Test set (s)</td>
</tr>
<tr>
<td>ResNet50</td>
<td>52</td>
<td>5.95</td>
</tr>
<tr>
<td>EfficientNetB5</td>
<td>90</td>
<td>15.58</td>
</tr>
<tr>
<td>VGG16</td>
<td>64</td>
<td>5.79</td>
</tr>
<tr>
<td>ResNext</td>
<td>78</td>
<td>8.15</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>34</td>
<td>4.46</td>
</tr>
</tbody>
</table>

Table 6. Running time comparison

Compared with the running time of common backbone image classification networks, the efficient AD prediction network proposed in this article is substantially shorter in both training time and testing time.

4.3 Ablation Experiments

To demonstrate that the adjustment of the convolutional network in this article is effective in enhancing the performance of network classification, the extended ablation experiment is subsequently designed in this article. First, the AD vs. NC classification experiments were conducted using networks designed with patches of different sizes. The effects were compared to those without dividing the patches. The experimental results are shown in Table 7.

<table>
<thead>
<tr>
<th>Patch Size</th>
<th>No Patch</th>
<th>4</th>
<th>7</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>97.29</td>
<td>98.33</td>
<td>98.33</td>
<td>98.54</td>
</tr>
<tr>
<td>F1-Score (%)</td>
<td>97.48</td>
<td>98.35</td>
<td>98.33</td>
<td>98.08</td>
</tr>
</tbody>
</table>

Table 7. Effect of different patch sizes on classification performance

By comparing network classification performance when using patches of different sizes, using the operation of partitioning the input image will significantly improve the classification performance. The classification accuracy will increase as the patch size becomes larger. The classification performance is best when the patch size is 14x14 pixels.

In addition, for the effect of convolution kernel size on the classification effect, the same ablation experiment is conducted (see Table 8).
The large size of the convolution kernel increases the network’s perception field. It also obtains a stronger classification performance. As shown in Table 7, the larger the size of the network convolution kernel, the higher the classification accuracy of the network. When increased to 7×7, the classification accuracy is improved to the maximum. It no longer improves.

In addition, this article draws ROC curves for the above two groups of ablation experiments respectively (see Figure 11).

The comparison of classification accuracy and ROC curves in the results of ablation experiments can prove that both the operation of dividing patch for input images and switching to large-size convolutional kernels are helpful to improve the classification performance of the AD network. The convolutional kernel size is increased to 7×7 performance saturation. To ensure the lightweight design of the network, the final network still chooses to use 7×7 convolutional kernels.

### 5. CONCLUSION AND FUTURE WORK

#### 5.1 Conclusion

This article migrates the segmentation patch operation and GELU activation function in the visual transformer to the convolutional network for application. The study reduces the size of the network and replaces the standard convolutional operation with depthwise separable convolution. The depthwise separable convolution separates the spatial and channel fusion operations of the feature map, reducing the single convolutional operation by separating the spatial and channel fusion operations of the feature map. The parameters of the proposed network is reduced to 1.3% as compared with other Alzheimer’s classification networks. A new convolutional neural network achieves high-performance classification, ensuring a lightweight model.
This article uses the brain MRI as a classification material, with all images from the ADNI dataset. The final performance (98.54%, 91.35%, 92.60%, and 93.33% accuracy) was achieved in the AD vs. NC, AD vs. MCI, MCI vs. NC, and AD vs. MCI vs. NC experiments. This outcome exceeded most of the previous 2D image Alzheimer’s classification methods. It even approaches some of the networks that use 3D imaging to assist in diagnosis.

The proposed novel Alzheimer’s classification network will provide a basis for physicians’ neuropsychological diagnosis of organic lesions. It will also reduce the dependence of the diagnostic effect on physicians’ skills, providing an Alzheimer’s diagnosis to more people. In addition, the comparison of experimental results shows that the classification accuracy of the network is not the highest among all existing works. Future studies should explore how to improve the classification accuracy of the network to promote a lightweight network.

5.2 Future Work

The classification and diagnosis methods of AD will diversify with the development of neurology. Some studies have tried multimodal diagnosis methods like the use of neuroimaging in conjunction with other medical indicators of the subjects to diagnose. Multimodal diagnosis could be the future direction of computer-aided medical diagnosis. In addition, PET imaging may increase in popularity, which can more accurately reflect brain metabolism and provide a strong basis for the diagnosis of AD. Future studies should explore multimodal diagnostic methods like the simultaneous use of PET imaging and MRI for the diagnosis of AD.

6. FUNDING

This work was supported by grants from the National Natural Science Foundation of China Project [No. 61972001], the Natural Science Foundation of Anhui Province Project [No. 1908085MF209]; and the National Undergraduate Training Program for Innovation and Entrepreneurship Project [No. 202110357089] [No. S202210357168].

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG02490) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
REFERENCES


Boan Ji is a senior undergraduate student at School of Computer Science and Technology, Anhui University, Hefei, China. His research interests are in pattern recognition and medical image processing.

Huabin Wang received a B.S. degree in computer science and technology from the Anhui University of Finance and Economics, China, in 2005, and a M.S. degree in signal and information processing, and a Ph.D. degree in computer application technology from Anhui University, Hefei, China, in 2008 and 2011, respectively. He is currently an Associate Professor and the Director of the School of Computer Science and Technology, Anhui University. His research interests include biometric recognition, template protection, and medical image processing.

MengXin Zhang is a senior undergraduate student at School of Computer Science and Technology, Anhui University, Hefei, China. Her research interests are in digital image processing and pattern recognition.

Borun Mao is a junior undergraduate student at School of Computer Science and Technology, Anhui University, Hefei, China. His research interests are in digital image processing and pattern recognition.

Xuejun Li received his Ph.D. degree in computer application technology from the School of Computer Science and Technology, Anhui University, Hefei, Anhui, China, in 2008. He is currently a full professor with the School of Computer Science and Technology, Anhui University, Hefei, Anhui, China. His major research interests include mobile edge computing, workflow systems, cloud computing, and intelligent software.