Alzheimer's disease detection using depthwise separable convolutional neural networks

Alzheimer's disease detection using depthwise separable convolutional neural networks

Junxiu Liu[1], Mingxing Li[1], Yuling Luo[1*], Su Yang[2], Wei Li[3], Yifei Bi[4]
[1] School of Electronic Engineering, Guangxi Normal University, Guilin, China.
[2] School of Computing and Engineering, University of West London, London, UK.
[3] Academy for Engineering and Technology, Fudan University, Shanghai, China.
[4] College of Foreign Languages, University of Shanghai for Science and Technology, Shanghai, China.
*yuling0616@gxnu.edu.cn

Abstract: To diagnose Alzheimer's disease (AD), neuroimaging methods such as magnetic resonance imaging have been employed. Recent progress in computer vision with deep learning (DL) has further inspired research focused on machine learning algorithms. However, a few limitations of these algorithms, such as the requirement for large number of training images and the necessity for powerful computers, still hinder the extensive usage of AD diagnosis based on machine learning. In addition, large number of training parameters and heavy computation make the DL systems difficult in integrating with mobile embedded devices, for example the mobile phones. For AD detection using DL, most of the current research solely focused on improving the classification performance, while few studies have been done to obtain a more compact model with less complexity and relatively high recognition accuracy. In order to solve this problem and improve the efficiency of the DL algorithm, a deep separable convolutional neural network model is proposed for AD classification in this paper. The depthwise separable convolution (DSC) is used in this work to replace the conventional convolution. Compared to the traditional neural networks, the parameters and computing cost of the proposed neural network are found greatly reduced. The parameters and computational costs of the proposed neural network are found to be significantly reduced compared with conventional neural networks. With its low power consumption, the proposed model is particularly suitable for embedding mobile devices. Experimental findings show that the DSC algorithm, based on the OASIS magnetic resonance imaging dataset, is very successful for AD detection. Moreover, transfer learning is employed in this work to improve model performance. Two trained models with complex networks, namely AlexNet and GoogLeNet, are used for transfer learning, with average classification rates of 91.40%, 93.02% and a less power consumption.

Keywords: Depthwise separable convolution, Alzheimer's disease, Deep learning, Transfer learning

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that can cause mental

disorders and even dementia in humans [1, 2]. AD patients are usually elderly, and a common symptom is the gradual loss of memory and understanding [3], which can inevitably lead to death. It is estimated that AD will suffer one in every 85 persons by 2050 [4]. So far, the exact cause of AD is still not quite clear. It has been reported that there are no effective medications or treatments that can prevent or reverse the progression of AD [5]. Therefore, it is critical to early diagnose the AD and design a treatment plan to slow the progression of AD. In recent years, the diagnosis of AD, especially its transitional phase, that is, mild cognitive impairment (MCI), has received growing attention [6]. Every year roughly 10 percent -15 percent of MCI patients are transitioned to AD [7]. It is found the converting from MCI to AD is often accompanied by the loss of the grey matter [8], abnormal changes in the volume of the medial temporal lobe structures [9], the decreased functional connectivity in the right superior frontal gyrus [10] and the decreased volumes of para hippocampal gyrus [11]. Based on these potential visual evidences of AD, the research approach should be developed that not only enhances the understanding of the pathophysiological processes of AD, but also contributes to the clinical study of AD.

Many neuroimaging techniques have been developed for exploiting the brain functions and structures, such as diffusion tensor imaging [12], magnetic resonance spectroscopy [13], electroencephalography [14], and magnetic resonance imaging (MRI) [15]. Recently, MRI has become increasingly popular in studying the brain nerve connections. MRI has shown tremendous promise as one type of well-developed brain imaging technology in providing detailed information for the diagnosis of high-level neurological disorders, such as depression and schizophrenia [16]. Rapid developments of neuroscience [17–22] and machine learning (ML) are widely used for automatic pattern recognition of clinical image data [23–26]. Recent studies have shown that in certain circumstances, ML algorithms can predict AD even better than clinicians [24], which is rather appealing and therefore, the computer-aided diagnosis has become an important research topic, due to its relatively low cost while training an expert system. Although statistical ML method such as support vector machine (SVM) [27] has shown some merits in automatic AD detection, a few recent deep learning (DL) methods have been found superior to the conventional statistical methods. The convolutional neural network (CNN) is popular in DL community thanks to its great success in image classification [28–30]. These achievements have attracted researchers to develop improved CNN-based systems for AD detection. However, despite great efforts have been made to improve the accuracy of classification, few works were done to optimize of the architecture of CNN for practical AD detection. In this paper, MRI-based feature is developed for AD classification using a depthwise separable convolution (DSC)-based CNN, and decent recognition accuracy rate is achieved.

The research approach of this work is divided in three-fold sequentially:

1) A CNN is designed to train and identify a small number of MRI with an

obtained high classification accuracy.

2) The CNN is further optimized to improve portability, which is a depthwise separable convolutional neural network. It decreases the number of parameters and the cost of computation, while the classification accuracy rate is maintained.

3) Two well-trained networks are used for transfer learning and good classification accuracies are achieved, which evidenced the effectiveness of the proposed depthwise separable convolutional neural network.

The remainder of the paper is structured as follows: similar works are summarized in Section II. Section III discusses the research methodology, and Section IV contains the experimental results and interpretation. Section V provides the conclusion and future work.

2. Related Works

The ML technique is commonly used in the automated pattern recognition based on images. [25], [31]. Classical ML algorithms, such as SVM algorithm [27] and linear judgment analysis algorithm [29], have been successfully applied to diagnose the early stage of AD using MRI. Recently, a feed-forward neural network [31], which used dual-tree complex wavelet transform for feature extraction, was proposed to classify the MRI. Detailed discussion and its comparisons with other popular methods were also addressed in [31]. A study on four-class classification was proposed in [32]: The study investigated the diagnosis of AD, late mild cognitive impairment (LMCI), early mild cognitive impairment (EMCI), and healthy control (HC). Multi-core SVM [33] and the weighted random SVM [34] have also been used for the same types of classification and the performance continuously improved. The detailed recognition accuracy rates from these works are shown in Table I.

Table I. Performance comparison of previous ML methods.

Methods	EMCI versus LMCI	LMCI versus AD
[32]	72.05%	81.70%
[33]	73.60%	90.10%
[34]	90.00%	88.89%

In the area of AD recognition, DL is often considered advantageous because it does not require complex feature engineering and generalization beforehand. Recently, DL methods have become increasingly popular, and arguably surpass the traditional methods. In [35], it was recommended that a flexible DL program implement dropout strategy to classify AD patients at various stages of development. The results indicate that the dropout has a good effect in the diagnosis of AD with its final average

classification rate reached 74.10%, improving the classification accuracy by 5.90% on average, compared to the classical DL methods. As one of DL's most used architectures, CNN has gained a great deal of interest in the area of image classification [30, 36]. An AD detection system based on CNN, AD patches and HC being used to train a CNN to recognize deep learning characteristics of MCI subjects, was introduced in [37], and the final accuracy of recognition exceeded 79.9% with 818 subjects.

A popular method [38] achieves a good classification by segmenting the entire brain into multiple anatomical or distinguished regions, and then extracting regional features. Another method [39] introduced that the features extracted from neuroimaging data are not isolated but have high correlations. Considering the relationship between these features, tree-guided sparse coding methods and resampling schemes using elastic nets have been proposed in [40] for AD diagnosis. The approach of [41] uses unsupervised CNN, PCANet, to achieve feature learning of MRI images. PCANet can learn the filters in CNN through traditional unsupervised machine learning algorithms. PCANet performs hash coding on the feature map obtained by the convolutional layer, and then uses histogram block coding, and finally outputs the extracted features. Although these DL algorithms provided good accuracy rates, the model structures are complex for deployments on the embedded devices with limited computing resources [18]. To address this challenge, this work aims to replace the standard convolution architecture of CNN by DSC to reduce the number of parameters and training time of the neural net model.

The fine-tuning of the networks based on transfer learning have also been explored using medical image data. Studies using medical images in [42, 43] shows that fine-tuning of the model based on transfer learning is better than training directly from scratch in most cases. Therefore, in this work the AlexNet and GoogLeNet models are separately used as the base for transfer learning to further classify AD. The results indicate the positive effectiveness of DSC for diagnosing AD.

3. Methodology

This section presents the methodology of the related methods proposed in this work, including CNN, DSC and transfer learning, as well as the pipeline of training and optimizing the neural network.

3.1. CNN model

As a multi-layer neural network, CNN is particularly effective when dealing with scenes involving a large number of images. The basic structure of a classical CNN consists of a convolutional layer, a pooling layer, and a fully connected layer. A classical CNN's basic structure consists of a convolutional layer, a pooling layer, and a totally fully layer. In detail, the convolutional layer is designed to extract different features of the image. The pooling layer further abstracts the original features, which greatly

reduces the training parameters and eases the over-fitting of the model. In summary, CNN allows a collection of features through the convolution kernel's filtering mechanism, which decreases the amount of network parameters through convolutional weight sharing and pooling activity. The soft-max classifier is inserted into the fully connected layer, after extracting the features, to classify the samples.

Figure 1 displays the overall CNN layout configuration and can be split into the module for extraction of functionality and the module for classification. 'Conv.' and 'Pool.' denote convolution operations and pooling operations, respectively.

For the feature extraction of this work, there are N gray-scale images $X_n, n \in [1, N]$ after the data pre-processing, and their pixels are scaled to a size of 56×56 and normalized to the interval [0, 1]. Moreover, the standard convolutional layer of the convolution kernel of size 3×3 is then fed for feature extraction. For each convolution operation, batch normalization (BN) [44] function and rectified liner unit (ReLU) activation function is implemented. Thereafter, each convolutional layer is accompanied by a maximum pooling of size 2×2, which samples down by half the previous feature map.

Three such standard convolutional layers are applied to this model. The CNN model framework is used as a benchmark in this work, as shown in Figure 1.

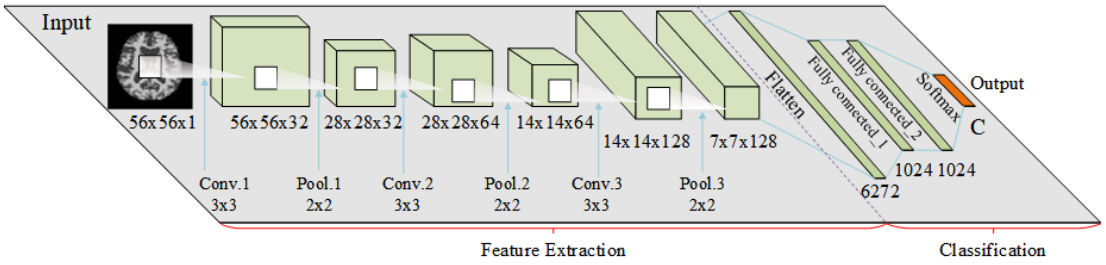

Figure 1. The overall design of the architecture CNN from this work. 'Conv.' and 'Pool.' denote the convolution and pooling processes, respectively.

A 7×7×128 feature matrix is fed to the classification module after the previous feature extraction. Firstly, the feature map is flattened to 6,272 feature vectors, and then the feature vectors are densified by using two fully connected layers, each layer is set to contain 1024 neurons. C is the number of classifications in AD dataset. Then, the C-dimensional score vector $S([S_1 \dots, S_l, \dots S_C])$ is expressed by the predictive probability with the soft-max function, and the value of each fraction is between [0, 1]. The soft-max function is given by

$$P(y_n = l|X_n) = \frac{exp\ (S_l)}{\sum_{l=1}^{C} exp\ (S_l)}, \quad (1)$$

where $P(y_n = l|X_n)$ is the forecasted likelihood for sample X_n to be class l.

To avoid over-fitting of the network, the popular dropout regularization is used for each pooling layer of the CNN model [45]: Some neurons in the neural network are discarded at random during model training. In this work, 10% of the neurons are randomly removed.

The network weight w and the cost function of the network need to be optimized during the process of CNN training. Regularized cross-entropy is used as cost-function in this analysis. The cost-function can be translated as

$$L(w) = \sum_{n=1}^{N} \sum_{l=1}^{C} y_{nc}\, log[P(y_n = l|X_n)] + \gamma l_2(w), \tag{2}$$

where y_{nc} is 0 if the X_n ground truth label is the lth dot, or if it is 1 otherwise. The l_2 regularization with its coefficient γ controls the weight w while training the model, also detects the limitation of the model space so that over-fitting may be avoided.

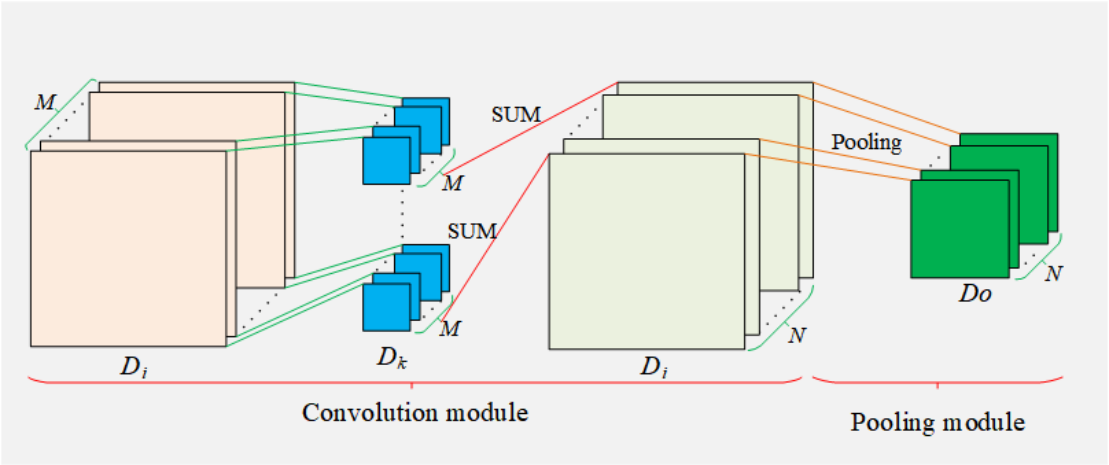

Figure 2. The structure of a standard convolution layer includes convolution module and pooling module. This framework is based on the case where the training step size is one and the input feature map is zero-padding.

Figure 2 shows the standard convolution process. A standard convolution layer takes a $D_i \times D_i \times M$ feature map I as input and generates a $D_i \times D_i \times N$ feature map output O, where D_i is the spatial width and height of the square input feature map, M is the number of input feature map channels, and N is the number of output feature map channels. Extracts function from the size $D_k \times D_k$ convolution kernel from the standard convolution layer. D_k is convolution kernel spatial width and height.

The standard convolution calculation process formula of the feature map I to the feature map O is given by

$$O_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot I_{k+i-1,l+j-1,m}, \tag{3}$$

where I represent the input features maps, O represents the output features maps, and k

represents the convolution kernels. i and j specify the Convolution kernel element location. k and l decide the location of the element in the input feature map and the output feature map, m represents the input feature map channel and n represents the output feature channel.

The parameters of standard convolution are computed as

$$F = M \times N \times {D_k}^2. \tag{4}$$

The computing cost of standard convolution is shown by

$$G = M \times N \times {D_i}^2 \times {D_k}^2, \tag{5}$$

where G represents the total number of parameters of the model, F represents the computational cost, M represents the number of channels of the input feature map, N represents the number of channels of the output feature map, D_i represents the spatial width and height squared input features of the object map, and D_k represents the convolution the spatial width and height of the convolution kernel.

3.2. DSC operation

The traditional convolution process uses weight sharing and pooling operations. Such techniques can significantly reduce the number of network parameters employed and the cost of computation, but still cannot satisfy the criteria of installing models on many embedded devices. In this work, A new approach for further reduce the number of parameters and the computational burden of a CNN is provided. The standard convolutional layer considers the input image data from the channel and space aspects simultaneously. DSC decomposes the traditional convolution into two sequential steps in order to reduce the potential redundancy of the standard convolution due to ignorance of information types: depthwise convolution followed by pointwise convolution (1×1 convolution kernels). In detail, DSC divides the standard convolutional layer into two layers, one for filtering and the other for extracting features with multiple 1×1 convolution kernel. Depthwise convolution first applies the convolution kernel to a channel of the image, and then the point-wise convolution is used to integrate the channel convolution output. The DSC uses a 1×1 convolution kernel instead of a 3×3 convolution kernel to process the input image data. It is found in such design the DSC greatly reduced model parameters and the computational complexity compared with standard convolution, and the experimental results will be analysed in detail.

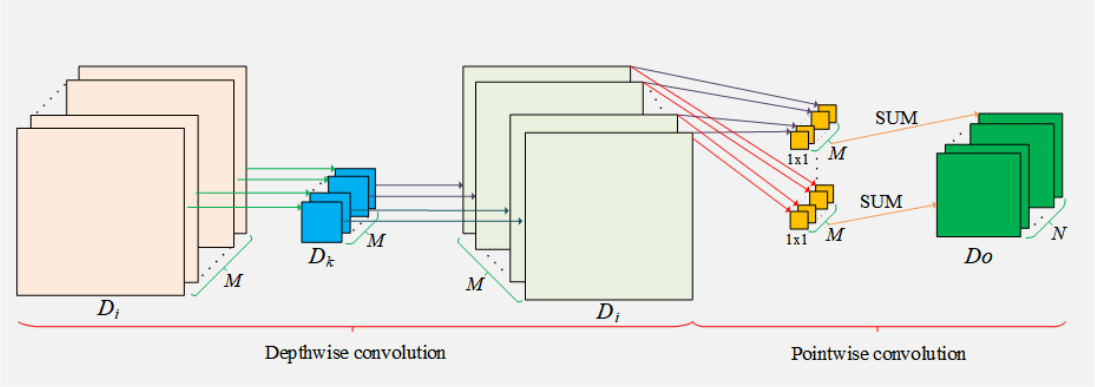

Figure 3. The structure of a DSC includes depthwise convolution module and pointwise convolution module. The stride is one and zero padding applies to depthwise convolution module.

The DSC structure is expressed in Figure 3. For each input channel, Depthwise convolution applies a single filter with the stride of one and zero padding. Pointwise convolution is then used to construct a linear combination of the depthwise convolution output with a convolution kernel of size 1×1. Pointwise convolution achieves the effect of down-sampling by adjusting the stride. This work uses BN [44] and ReLU nonlinear function for both DSC layers.

The feature map for the output of the depthwise convolution is expressed as

$$\bar{O}_{k,l,m} = \sum_{i,j} K_{i,j,m} . I_{k+i-1,l+j-1,m}, \tag{6}$$

where I represent the input feature maps, $\bar{O}$ represents the output feature maps, and K represents the convolution kernels. i and j determine the element position of the convolution kernel. k and l decide the location of the input feature map element and the output feature map, m represents the input feature map channel.

The parameter calculation and cost function for the depthwise convolution can be denoted by

$$F_2 = M \times {D_K}^2, \tag{7}$$

and

$$G_2 = M \times {D_i}^2 \times {D_K}^2. \tag{8}$$

The number of parameters is related only to the number of input feature mapping channels and the convolution kernels. The computational cost is proportional to the number of input feature mapping sources, the convolution kernel and the square input feature mapping function. The parameters and computing costs of depthwise convolution do not need to consider the output feature mapping N. Compare to formulas (4) and (5), the formulas (7) and (8) above clearly demonstrate the simplicity of the

depthwise convolution. However, unlike the conventional convolution layer, DSC only filters input channels without combining them into new features. Therefore, this paper attempts to merge the performance features of the depthwise convolution layer with the pointwise convolution in order to produce new features.

The parameter formula for DSC is calculated by

$$F_3 = M \times {D_K}^2 + M \times N. \tag{9}$$

The calculation cost formula for DSC is given by

$$G_3 = M \times {D_i}^2 \times {D_K}^2 + M \times N \times D_i^2. \tag{10}$$

DSC based on 3×3 convolution kernel is used in this work, which computes eight to nine times faster than the standard convolution, achieved a comparable accuracy (shown in Section V).

The parameter reduction is described by

$$F_4 = F_3 - F = M \times D_k^2 + M \times N - M \times N \times D_k^2. \tag{11}$$

The calculation cost reduction is given by

$$G_4 = G_3 - G = M \times D_i^2 \times D_k^2 + M \times N \times D_i^2 - M \times N \times D_i^2 \times D_k^2. \tag{12}$$

When training a neural network, BN function, ReLU function and pooling layer are used after each standard convolution layer. In DSC, BN and ReLU function are used. Their structure is shown in Figure 4.

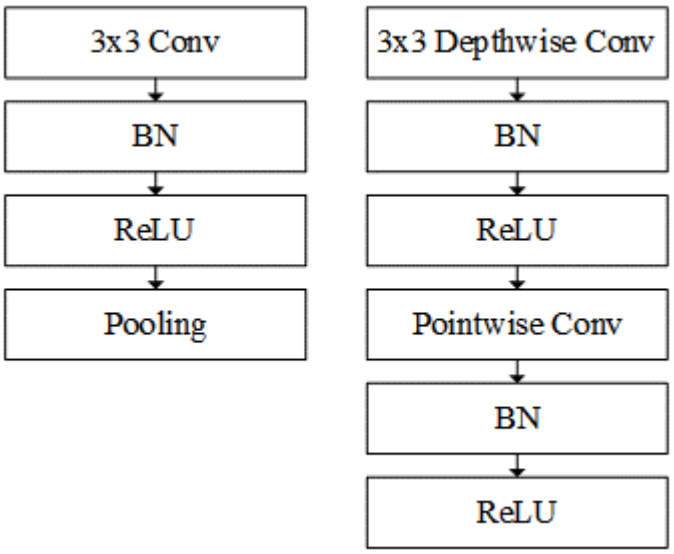

Figure 4.Standard convolution with BN, ReLU and pooling layer (Left), and DSC with depthwise and pointwise layers followed by BN and ReLU (Right).

Specially, the standard convolution feature map is down-sampled by the pooling

layer, and the down-sampling in the DSC is achieved by adjusting the convolution stride.

Table II. DSC network Architecture

Type / Stride	Filter Shape	Input Size
Conv / s1	3×3×1×32	56×56×1
Conv dw / s2	3×3×32	56×56×32
Conv / s1	1×1×32×64	28×28×32
Conv dw / s2	3×3×64	28×28×64
Conv / s1	1×1×64×128	14×14×64
Avg Pool / s1	Pool 2×2	14×14×128
FC_1 / s1	6272×1024	1×1×6272
FC_2 / s1	1024×3	1×1×1024
SoftMax / s1	classifier	1×1×3

Table II shows a body architecture of the DSC used in this work. This architecture is an optimization of the previous standard convolution architecture, replacing the two standard convolutional layers of the standard convolutional architecture with two DSC layers. The pooling module in the standard convolutional layer performs a size 2×2 down sampling operation on the input feature map. In pointwise convolution of DSC, the stride is set to two, which can effectively achieve the down-sampling operation, *s1* means the stride of convolution is one, and *s2* means the stride of convolution is two.

3.3. Transfer learning

For small data sets, the classification accuracy rate would be relatively low if CNN are trained from the scratch by back propagation. In order to leverage multiple pre-trained networks, it is possible to obtain a higher classification accuracy rate through transfer learning. In transfer learning, the network model uses pre-trained network. Its weights are pre-set, and only the last fully connected layer is retrained. In this work, two popular architectures are used including:

1). AlexNet: AlexNet was proposed in [46] and won the 2012 ILSVRC competition. The top5 error rate was 16.4%, the second-best contest entry was 26.2% error rate. AlexNet's network structure contains eight neural networks, including five convolutional layers and three fully connected layers, containing 630 million links, 650,000 neurons and 60 million parameters.

2). GoogLeNet: GoogLeNet, a new DL structure, was proposed in [47], which won the ILSVRC championship in 2014 and reduced the error rate of Top5 to 6.67%.

GoogLeNet uses 22 layers of neural networks, but the parameters are only half that of AlexNet. Google LeNet points out that the best way to achieve high-quality models is to increase the model depth, but wider networks are vulnerable to overfitting and computational complexity. GoogLeNet converts some convolution and fully connection into a sparse connection, and propose for this reason a modular system called Inception.

4. Results

The dataset used in this work is first presented in this section, then the results of CNN, DSC and transfer learning algorithms on AD detection are analysed.

A series of comparisons are presented: the results of CNN are compared with other relevant algorithms. The results of DSC are further compared with the standard CNN algorithm. Finally, the results of transfer learning are analysed.

4.1. Dataset

In this paper the Open Access Sequence of Image Studies (OASIS) structural MRI data is used. [48]. OASIS is a project that is intended to provide the scientific community free access to brain neuroimaging datasets. The examples from HC, MCI and AD groups are shown in Figure 5. OASIS provides two types of data, cross-sectional data and longitudinal data. Because the purpose of this paper is to classify data sets into two and three categories, cross-sectional data meets the requirements. The data collection contains a cross-sectional sample of 416 subjects aged 18 to 96. 3 to 4 separate T1-weighted MRI scans are obtained from a single scan for each subject. The subjects are right-handed, including men and women. Clinically, 100 subjects over the age of 60 had been diagnosed with very mild to moderate AD, among them. Among them, 100 subjects over 60 years old had been clinically diagnosed with very mild to moderate AD. In addition, the reliability dataset, which contains 20 non-dementia subjects, was tested again 90 days after their initial meeting. Whether the subjects in the dataset were ill was determined by the clinical dementia rating (CDR) variable, ranging from zero to two. Hypothesis zero represents HC, two represents AD, and the rest are MCI.

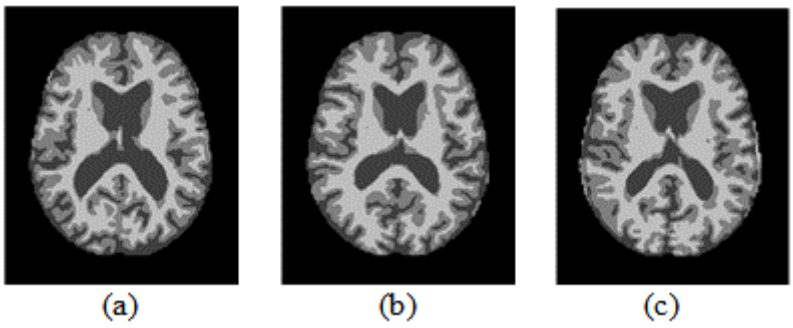

Figure 5. Images from the OASIS MRI dataset (a) HC. (b) MCI. (c) AD.

The dataset includes 332 HC, 68 MCI and 30 patients with severe AD. Data of patients with MCI and AD are over-sampled, which can expand the amount of data and avoid the impact of data imbalance. At the same time, HC data is under sampled. After resampling, the final dataset includes 266 HC images, 136 MCI images and 90 images of patients with severe AD. Finally, data enhancement processing (clipping, flipping, increase contrast, rotate etc.) are performed on the OASIS dataset. After data enhancement, 532 HC images, 408 MCI images and 450 images of patients with severe AD were obtained. During training, data enhancement is also performed on the model, e.g. dropout [45] technology is applied. The OASIS dataset is randomly broken down into five sections, where cross-validation is used five times during training. The experiments in this paper secured that the patient-wise division is taken into account.

In this paper, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is used as a test set to test the performance of the model. The ADNI is a longitudinal multiple centers study aimed at the development of clinical, imaging, genetic and biochemical biomarkers, as well as early detection and tracking of AD. At each stage of the ADNI dataset, new participants were recruited across North America and agreed to complete various imaging and clinical evaluations. This has made a significant contribution to AD the research.

4.2. Experimental results for CNN algorithm

The CNN model is trained for two classification scenarios in order to test the efficiency of the CNN model developed in this paper: binary and three-class classifications. The training for binary classification is conducted for two cases, i.e. HC versus MCI, and MCI versus AD. The demonstrate proposed in this paper is compared to other models, and a better performance is achieved. The classification accuracy rate of health control and mild cognitive impairment reach 84.65%, and the classification accuracy rate of mild cognitive impairment and AD is 72.96%. A comparison with other methods is shown in Table III.

When classifying HC and MCI, training 100 epochs, the training loss and verification loss can converge quickly, the final training loss can reach 0.3919, and the verification loss can reach 0.4048. Their convergences are shown in Figure 6. When classifying MCI and HC, training 100 epochs, the training loss and verification loss can converge quickly, the final training loss can reach 0.4062, and the verification loss can reach 0.4243. Their convergences are shown in Figure 7. The training loss of HC and MCI is 0.0143 which is lower than MCI and AD, and the verification loss of HC and MCI is 0.0195 which is lower than MCI and AD. It can be seen from the experiment that because the number of samples of HC is larger, the training loss and verification loss of HC and MCI are lower.

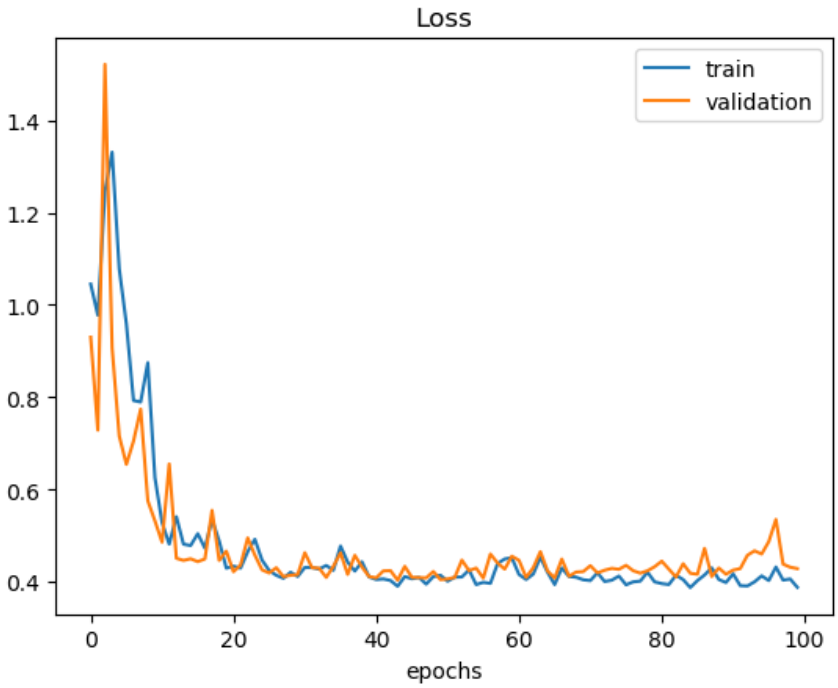

Figure 6. Training loss and validation loss for classifying HC and MCI.

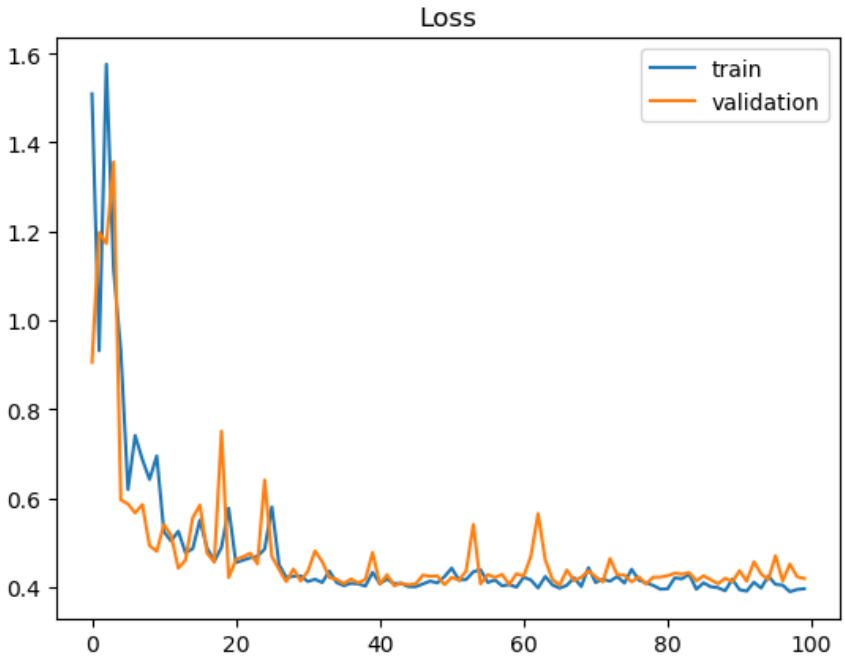

Figure 7. Training loss and validation loss for classifying MCI and AD.

In Table III, ACC stands for accuracy, SEN stands for sensitivity, SPC stands for specificity, and AUC stands for area under curve. EMCI describes a mild cognitive disability at an early stage and LMCI is a mild late cognitive impairment. Among all

these methods, the classification accuracy rate obtained from the proposed method appears to be outstanding. The sample sizes of the different classes of dataset used in this work are quite different compared to the datasets used in other methods. Among patients with MCI and AD, the number of samples is small. In particular, the sample size of HC and AD differs greatly, which is the main reason for the low classification accuracy rate between them. In 266 HC samples and 136 MCI samples, the classification accuracy rate is 84.65%. In addition, HC, MCI and AD are also classified into three classes, and the classification accuracy rate is 78.02%. Compared with other advanced methods, the proposed CNN method has better classification performance. For the proposed method, the sensitivity is 83.21%, the specificity is 82.15%, and the AUC is 85.23%. Compared with other methods, the proposed method achieves a similar detection performance. There are some minor differences under several specific metrics which is mainly due to that the datasets used in the approaches and the number of samples are different. However the main advantages of the proposed method are efficient network design and significantly reduced parameters and more details will be provided in next subsection. In particular, this work uses the ADNI dataset to test the generalization of the model. 353 MCI and 99 AD images are selected by ADNI dataset. ACC reaches 75.32%, SEN reaches 80.13%, SPC reaches 65.32%, and AUC reaches 85.23%. It can be seen from the test results of ADNI data that there is a difference in SPC, which may be due to the differences between OASIS and ADNI.

Table III. Performance comparison with other methods.

Methods	Number of samples	ACC	SEN	SPC	AUC
Random forest [32]	164 EMCI versus 189 LMCI	72.50%	79.00%	68.70%	78.50%
	189 LMCI versus 99 AD	81.70%	83.50%	72.80%	84.30%
Random forest [25]	229 HC versus 188 AD	75.00%	72.00%	64.00%	-
CNN [37]	229 HC versus 188 AD	79.90%	84.00%	74.80%	86.10%
Multi-kernel SVM [33]	114 EMCI versus 91 LMCI	78.80%	74.40%	82.10%	78.30%
SVM [49]	33 HC versus 57 AD	81.10%	60.60%	93.00%	-
SVM [50]	127 HC versus 67 MCI	74.90%	61.10%	83.40%	-
Regression analyses [51]	42 HC versus 38 AD	82.50%	-	-	-
This work	266 HC versus 136 MCI	84.65%	82.35%	79.50%	85.23%
	136 MCI versus 90 AD	72.96%	78.34%	82.15%	77.56%
	226 HC, 136 MCI versus 90 AD	78.02%	83.21%	75.32%	83.45%

353 MCI versus 99 AD (ADNI)	75.32%	80.13%	65.32%	81.41%

4.3. Analysis for DSC algorithm

The DSC is used to optimize the CNN model. The CNN model in this paper uses three standard convolutions and they are replaced by DSCs. Comparing to the optimized depthwise separable models with the standard CNN model, the results are shown in Table IV.

Table IV. Depthwise separation convolution VS standard convolution.

Resolution	ACC	AUC	Million Mult-Adds	Thousand Parameters
CNN model	78.02%	83.45%	29.804544	92.448
One DSC	77.91%	83.23%	29.029952	92.201
Two DSC	77.85%	82.35%	16.410688	76.105
Three DSC	77.79%	81.95%	3.678528	11.145

In Table IV, the levels of classification accuracy, the number of parameters and computational costs of the optimized DSC neural network are compared with the standard convolution model. The conventional convolutions are replaced by deep separable convolutions with less computing cost and model parameters, and very close classification accuracy and AUC. In particular, when all three standard convolution layers of the complete standard convolution model are replaced by DSC, the classification accuracy rate only decreases by 0.23% and the ACC rate decreases by 1.50%, whereas the advantage of the proposed neural network reduces the number of million mult-adds by 84.25% and the thousand parameters by 87.94%.

In the meantime, the advantage of the proposed model is that the number of million mult-adds is significantly reduced by 84.25%, and the thousand parameters is reduced by 87.94%.

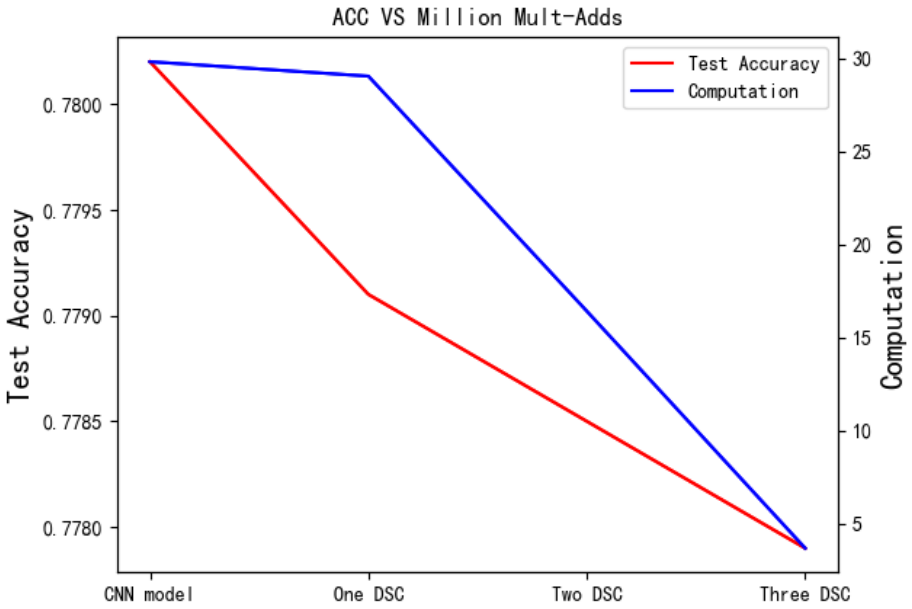

Figure 8. When testing with the OASIS dataset, there is a trade-off between the computational cost of the model and the accuracy of the test.

Figure 8 shows the relationship between test accuracy rate and computing cost between the CNN model and the DSC model. After a standard convolution layer is replaced by the DSC layer, it is found that the test accuracy rate is reduced by 0.11%, but the computing cost is reduced by 40.07%. After two standard convolution layers is replaced with DSC layers, it is found that the test accuracy rate is reduced by 0.17%, but the computing cost is reduced by 81.65%. When all three standard convolution layers of the complete standard convolution model are replaced by DSC layers, the classification accuracy rate only decreases by 0.23%. In the meantime, the advantage of the proposed model is that computing cost is reduced by 84.25%. Therefore, as more convolutional layers are replaced, the computing cost decreases, and the model proposed achieves a reasonable trade-off between accuracy and the computational cost.

Figure 9 shows the relationship between test accuracy rate and model parameter between the CNN model and the DSC model. After a standard convolution layer is replaced by the DSC layer, it is found that the test accuracy rate is reduced by 0.11%, but the model parameter is reduced by 16.99%. After two standard convolution layers is replaced with DSC layers, it is found that the test accuracy rate is reduced by 0.17%, but the model parameter is reduced by 87.68%. When all three standard convolution layers of the complete standard convolution model are replaced by DSC layers, the classification accuracy rate only decreases by 0.23%. In the meantime, the advantage of the proposed model is that model parameter is reduced by 87.94%. Therefore, as more convolutional layers are replaced, the model parameter decreases, and the model proposed achieves a reasonable trade-off between model parameter and accuracy.

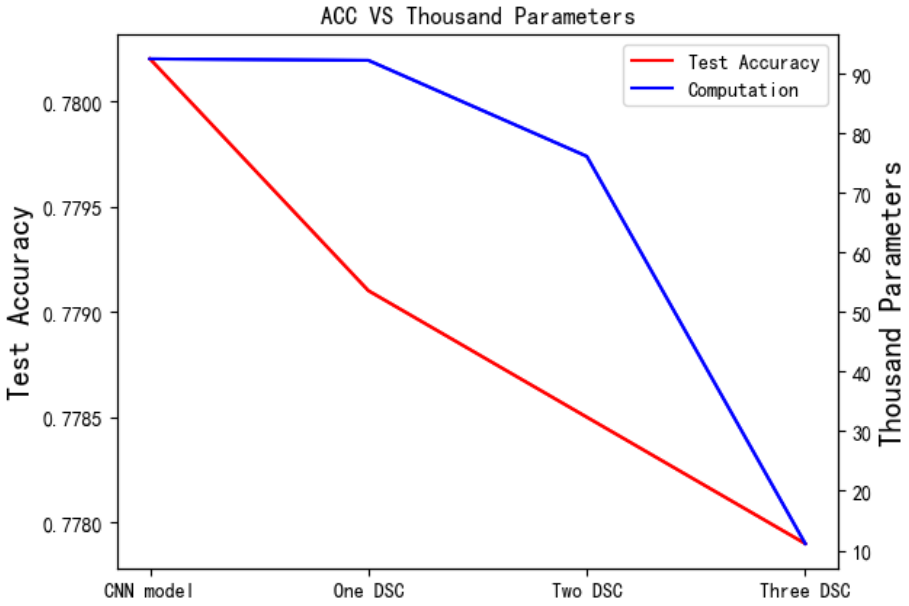

Figure 9. When testing with the OASIS dataset, there is a trade-off between the parameter of the model and the accuracy of the test.

Comparative experiments show that the DSC layer is used to replace more conventional convolution layers in the CNN model, and this achieves a lower computational cost and parameters of the model while maintaining the test accuracy. Therefore, the AD classification system using the DSC model is very beneficial for embedded devices with limited computing resources. The proposed method in this work has low computational cost and low number of parameters, but the generalization performance can be further investigated. This can be addressed by optimizing the proposed model in the future work.

4.4. Results of transfer learning algorithm

In the case of small dataset, this may lead to over-fitting or under-fitting, and the classification accuracy of training a neural network model from scratch is generally not high. The pre-trained model is used for transfer learning, and the rate of accuracy is substantially increased. The models AlexNet and GoogLeNet are pre-trained on the ImageNet dataset, and are then used for transfer learning in this work.

Table V shows the accuracy rate results for the four models, all of which are trained using the OASIS MRI dataset. It can be seen from Table V that due to insufficient training data, the classification accuracy of CNN and DSC models trained from scratch is low. The pre-trained AlexNet and GoogLeNet models are fine-tuned using the OASIS MRI data, and the classification accuracy rates are significantly improved. The pre-training models of AlexNet and GoogLeNet are based on ImageNet data which includes a large amount of data, so they have very good generalization ability and can achieve a

good performance when applied to OASIS MRI data. The AlexNet and GoogLeNet models obtain classification accuracy rates of 91.40% and 93.02%, respectively. GoogLeNet uses more convolutions and deeper layers than AlexNet, so classification accuracy is higher. Note that both the AlexNet model and the GoogLeNet model use 5-fold cross-validation and 500 iterations of training during transfer learning.

Table V. Test models and corresponding average accuracy rates.

Model	Acc. (%)
CNN (from scratch)	78.02
DSC (from scratch)	77.79
AlexNet (transfer learning)	91.40
GoogLeNet (transfer learning)	93.02

In Table V, it can be seen that transfer learning can achieve higher classification results, but AlexNet and GoogLeNet are very complex neural networks, and their computations are very intensive. Moreover, their frameworks also contain many standard convolution modules which can be replaced by the proposed DSC module to reduce the network complexity. This is one option for future research work.

5. Conclusion

A novel DSC network-based method for detection of AD is proposed in this paper. The conventional CNN method is first used to detect AD, and the classification accuracy rate reached 78.02% in a three-way classification scenario (AD, MCI and normal). Then, an AD detection method combining DSC and CNN is proposed. Compared with the CNN, the model parameters of the proposed method are reduced by 87.94% and the computing cost is reduced by 84.25%, where the classification accuracy rate remains moderately the same. It is quite promising in implementing the functionality of AD detection on mobile embedded devices with limited computing resources. Experiments on OASIS MRI dataset show that DSC method has great potential for AD recognition. The common training models of AlexNet and GoogLeNet are used for transfer learning to improve the classification accuracy rate of AD detection, and a good result is obtained in this paper. Consequently, one potential future work will consider combining DSC with AlexNet or GoogleNet to further increase the AD classification accuracy rate and to obtain a more compact model. At the same time, using the proposed method in other application areas can also be investigated in the future.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under Grants 61976063, the funding of Overseas 100 Talents Program of Guangxi Higher Education.

References

[1] T. Tong, Q. Gao, R. Guerrero, C. Ledig, L. Chen, and D. Rueckert, "A novel grading biomarker for the prediction of conversion from mild cognitive impairment to alzheimer's disease.," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 1, pp. 155–165, 2017.

[2] E. J. Kodis, S. Choi, E. Swanson, G. Ferreira, and G. S. Bloom, "N-methyl-D-aspartate receptor-mediated calcium influx connects amyloid-β oligomers to ectopic neuronal cell cycle reentry in alzheimer's disease," *Alzheimer's Dement.*, vol. 14, no. 10, pp. 1302–1312, 2018.

[3] D. S. Roy, A. Arons, T. I. Mitchell, M. Pignatelli, T. J. Ryan, and S. Tonegawa, "Memory retrieval by activating engram cells in mouse models of early Alzheimer's disease," *Nature*, vol. 531, no. 7595, pp. 508–512, 2016.

[4] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of alzheimer's disease," *Alzheimer's Dement.*, vol. 3, no. 3, pp. 186–191, 2007.

[5] J. E. Morley, S. A. Farr, and A. D. Nguyen, "Alzheimer Disease," *Clin. Geriatr. Med.*, vol. 14, no. 3, pp. 129–135, 2018.

[6] X. Cui *et al.*, "Classification of alzheimer's disease, mild cognitive impairment, and normal controls with subnetwork selection and graph kernel principal component nnalysis based on minimum spanning tree brain functional network," *Front. Comput. Neurosci.*, vol. 12, no. 5, pp. 1–12, 2018.

[7] M. Grundman, P. RC, F. SH, and E. Al, "Mild cognitive impairment can be distinguished from alzheimer disease and normal aging for clinical trials," *Arch. Neurol.*, vol. 61, no. 1, pp. 59–66, 2004.

[8] G. B. Karas *et al.*, "Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease," *Neuroimage*, vol. 23, no. 2, pp. 708–716, 2004.

[9] L. Younes, M. Albert, and M. I. Miller, "Inferring changepoint times of medial temporal lobe morphometric change in preclinical alzheimer's disease," *NeuroImage Clin.*, vol. 5, pp. 178–187, 2014.

[10] E.-S. Lee *et al.*, "Default mode network functional connectivity in early and late mild cognitive impairment," *Alzheimer Dis. Assoc. Disord.*, vol. 30, no. 4, pp. 289–296, 2016.

[11] C. Echávarri *et al.*, "Atrophy in the parahippocampal gyrus as an early biomarker of alzheimer's disease," *Brain Struct. Funct.*, vol. 215, no. 3–4, pp.

265–271, 2011.

[12] G. Coppola *et al.*, “Altered thalamic microstructure in migraine without aura patients: a diffusion tensor magnetic resonance imaging study,” *Eur. J. Neurol.*, vol. 21, no. 2, pp. 287–292, 2014.

[13] A. Busato, P. F. Feruglio, P. P. Parnigotto, P. Marzola, and A. Sbarbati, “In Vivo imaging techniques: a new era for histochemical analysis,” *Eur. J. Histochem.*, vol. 60, no. 4, pp. 273–279, 2016.

[14] O. Sipilä *et al.*, “Transmission imaging for registration of ictal and interictal single-photon emission tomography, magnetic resonance imaging and electroencephalography,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 27, no. 2, pp. 202–205, 2000.

[15] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults,” *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2677–2684, 2010.

[16] M. J. Rosa *et al.*, “Sparse network-based models for patient classification using fMRI,” *Neuroimage*, vol. 105, no. 1, pp. 493–506, 2015.

[17] Y. Luo, L. Wan, J. Liu, J. Harkin, and Y. Cao, “An efficient, low-cost routing architecture for spiking neural network hardware implementations,” *Neural Process. Lett.*, vol. 48, no. 3, pp. 1777–1788, 2018.

[18] J. Liu, M. Li, Y. Luo, S. Yang, and S. Qiu, “Human body posture recognition using wearable devices,” in *International Conference on Artificial Neural Networks (ICANN)*, 2019, pp. 326–337.

[19] J. Liu, J. Zhang, Y. Luo, S. Yang, J. Wang, and Q. Fu, “Mass spectral substance detections using long short-term memory networks,” *IEEE Access*, vol. 7, no. 1, pp. 10734–10744, 2019.

[20] L. Wan, J. Liu, J. Harkin, L. McDaid, and Y. Luo, “Layered tile architecture for efficient hardware spiking neural networks,” *Microprocess. Microsyst.*, vol. 53, no. 1, pp. 21–32, 2017.

[21] Y. Luo *et al.*, “Low cost interconnected architecture for the hardware spiking neural networks,” *Front. Neurosci.*, vol. 12, no. 10, pp. 1–14, 2018.

[22] J. Liu, Y. Huang, Y. Luo, J. Harkin, and L. McDaid, “Bio-inspired fault detection circuits based on synapse and spiking neuron models,” *Neurocomputing*, vol. 331, no. 1, pp. 473–482, 2019.

[23] S. Yang, J. M. sanchez Bornot, K. Wong-Lin, and G. Prasad, “M/EEG-based bio-markers to predict the mild cognitive impairment and alzheimer’s disease: a aeview from the machine learning perspective,” *IEEE Trans. Biomed. Eng.*, vol. 1, no. 1, pp. 1–12, 2019.

[24] S. Klöppel *et al.*, “Accuracy of dementia diagnosis - a direct comparison between radiologists and a computerized method,” *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[25] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, “Machine learning framework for early mri-based alzheimer’s conversion prediction in mci subjects,” *Neuroimage*, vol. 104, no. 1, pp. 398–412, 2015.

[26] P. Cao, X. Shan, D. Zhao, M. Huang, and O. Zaiane, “Sparse shared structure based multi-task learning for MRI based cognitive performance prediction of alzheimer’s disease,” *Pattern Recognit.*, vol. 72, no. 12, pp. 219–235, 2017.

[27] C. Plant *et al.*, “Automated detection of brain atrophy patterns based on MRI for the prediction of alzheimer’s disease,” *Neuroimage*, vol. 50, no. 1, pp. 162–174, 2010.

[28] A. Dos Santos Siqueira, C. E. Biazoli Junior, W. E. Comfort, L. A. Rohde, and J. R. Sato, “Abnormal functional resting-state networks in ADHD: graph theory and pattern recognition analysis of fMRI data,” *Biomed Res. Int.*, vol. 2014, no. 1, pp. 1–10, 2014.

[29] M. Goryawala *et al.*, “Inclusion of neuropsychological scores in atrophy models improves diagnostic classification of Alzheimer’s disease and mild cognitive impairment,” *Comput. Intell. Neurosci.*, vol. 2015, no. 1, pp. 1–14, 2015.

[30] X. Jia, X. Xu, B. Cai, and K. Guo, “Single image super-resolution using multi-scale convolutional neural network,” *Sensors*, vol. 18, no. 3, pp. 1–17, 2018.

[31] D. Jha, J. I. Kim, and G. R. Kwon, “Diagnosis of alzheimer’s disease using dual-tree complex wavelet transform, pca, and feed-forward neural network,” *J. Healthc. Eng.*, vol. 2017, no. 5, pp. 1–13, 2017.

[32] S. H. Nozadi and S. Kadoury, “Classification of alzheimer’s and mci patients from semantically parcelled pet images: a comparison between av45 and fdg-pet,” *Int. J. Biomed. Imaging*, vol. 2018, no. 1, pp. 1–13, 2018.

[33] B. Jie, M. Liu, and D. Shen, “Integration of temporal and spatial properties of dynamic connectivity networks for automatic diagnosis of brain disease,” *Med. Image Anal.*, vol. 47, no. 1, pp. 81–94, 2018.

[34] X. Bi, Q. Xu, X. Luo, Q. Sun, and Z. Wang, “Analysis of progression toward alzheimer’s disease based on evolutionary weighted random support vector machine cluster,” *Front. Neurosci.*, vol. 12, no. October, pp. 1–11, 2018.

[35] F. Li, L. Tran, K. H. Thung, S. Ji, D. Shen, and J. Li, “A Robust Deep Model for Improved Classification of AD/MCI Patients,” *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 5, pp. 1610–1616, 2015.

[36] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum, and J. Mongan, “High-

throughput classification of radiographs using deep convolutional neural networks," *J. Digit. Imaging*, vol. 30, no. 1, pp. 95–101, 2017.

[37] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, and Y. Yang, "Convolutional neural networks-based MRI image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Front. Neurosci.*, vol. 12, no. November, pp. 1–13, 2018.

[38] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease", Neuroimage., vol. 1, no. 65, pp.167–175, 2012.

[39] C. Chu, A.L. Hsu, K.H. Chou, P. Bandettini, C. Lin, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images", Neuroimage., vol. 1, no. 60, pp. 1–13, 2012.

[40] E. Janoušová, M. Vounou, R. Wolz, K.R. Gray, D. Rueckert, G. Montana, "Biomarker discovery for sparse classification of brain images in Alzheimer's disease", Annals of the BMVA., vol. 2012, no. 2, pp. 1–11, 2012.

[41] Bi, X., Li, S., Xiao, B., Li, Y., Wang, G., & Ma, X. "Computer Aided Alzheimer's Disease Diagnosis by An Unsupervised Deep Learning Technology", Neurocomputing. vol. 1, no. 1, pp. 1–9, 2019

[42] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: full training or fine tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[43] M. Hon and N. Khan, "Towards alzheimer's disease classification through transfer learning," in *Bioinformatics and Biomedicine(BIBM)*, 2017, pp. 1166–1169.

[44] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (PMLR)*, 2015, pp. 448–456.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] G. Antonellis *et al.*, "ImageNet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1097–1105.

[47] G. Zeng, Y. He, Z. Yu, X. Yang, R. Yang, and L. Zhang, "Going deeper with convolutions christian," *Comput. Vis. Pattern Recognit.*, vol. 91, no. 8, pp. 1–9, 2015.

[48] A. F. Fotenos, A. Z. Snyder, L. E. Girton, J. C. Morris, and R. L. Buckner, "Normative estimates of cross-sectional and longitudinal brain volume decline

in aging and AD," *Neurology*, vol. 64, no. 6, pp. 1032–1039, 2011.

[49] S. Klo *et al.*, "Automatic classification of MR scans in alzheimer's disease," *Bain*, vol. 131, no. 3, pp. 681–689, 2008.

[50] Q. Zhou, M. Goryawala, M. Cabrerizo, J. Wang, W. Barker, and D. Loewenstein, "An optimal decisional space for the classification of alzheimer's disease and mild cognitive impairment," *IEEE Trans. Biomed. Eng.*, vol. 6, no. 8, pp. 2245–2253, 2014.

[51] J. Brewer, L. K. Mcevoy, R. G. Jennings, D. Karow, and A. M. Dale, "Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of alzheimer disease," *Am. J. Neuroradiol.*, vol. 31, no. 2, pp. 347–354, 2010.