Journal of Medical Systems (2018) 42: 85
https://doi.org/10.1007/s10916-018-0932-7

IMAGE & SIGNAL PROCESSING

# Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling

Shui-Hua Wang[1,2] · Preetha Phillips[3] · Yuxiu Sui[4] · Bin Liu[5] · Ming Yang[6] · Hong Cheng[7]

Received: 29 December 2017 / Accepted: 6 March 2018 / Published online: 26 March 2018

## Abstract

Alzheimer's disease (AD) is a progressive brain disease. The goal of this study is to provide a new computer-vision based technique to detect it in an efficient way. The brain-imaging data of 98 AD patients and 98 healthy controls was collected using data augmentation method. Then, convolutional neural network (CNN) was used, CNN is the most successful tool in deep learning. An 8-layer CNN was created with optimal structure obtained by experiences. Three activation functions (AFs): sigmoid, rectified linear unit (ReLU), and leaky ReLU. The three pooling-functions were also tested: average pooling, max pooling, and stochastic pooling. The numerical experiments demonstrated that leaky ReLU and max pooling gave the greatest result in terms of performance. It achieved a sensitivity of 97.96%, a specificity of 97.35%, and an accuracy of 97.65%, respectively. In addition, the proposed approach was compared with eight state-of-the-art approaches. The method increased the classification accuracy by approximately 5% compared to state-of-the-art methods.

**Keywords** Alzheimer's disease · Convolutional neural network · Leaky rectified linear unit · Max pooling · Data augmentation · Activation function

## Background

Alzheimer's disease (AD) is a progressive brain disease. AD eventually leads to the death of nerve cells within the brain [1], and the brain volume will shrink. It was estimated that one out of 85 people would be affected by AD by 2050. Several popular non-invasive neuroimaging tools are used to study AD, such as positron emission tomography (PET) [2] and magnetic resonance imaging (MRI) [3] MRI is one of the most popular methods, since it can provide a good resolution of soft tissues within the brain.

Many MRI-based classification methods were proposed during the last decade, using various computer vision approaches. The goal is to assist neuro-radiologists to make decisions whether AD or healthy, on the basis of brain images. For example, Plant, Teipel [4] combined brain region cluster (BRC) and Bayes method. Their average accuracy reached 92.00%. Savio and Grana [5] when using the trace of Jacobian matrix (TJM) method, and their average accuracy achieved 92.83 ± 0.91%. Gray, Aljabar [6] utilized random forest (RF) approach. Zhang [7] combined displacement field (DF) with support vector machine (SVM). Wang [8] proposed

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Shui-Hua Wang
shuihuawang@ieee.org

✉ Preetha Phillips
pphillips@osteo.wvsom.edu

✉ Hong Cheng
ch8706@sohu.com

[1] Department of Informatics, University of Leicester, Leicester LE1 7RH, UK

[2] Department of Electrical Engineering, The City College of New York, CUNY, New York, NY 10031, USA

[3] West Virginia School of Osteopathic Medicine, 400 N Lee St, Lewisburg, WV 24901, USA

[4] Department of Psychiatry, Affiliated Nanjing Brain Hospital of Nanjing Medical University, Nanjing, People's Republic of China

[5] Department of Radiology, Zhong-Da Hospital of Southeast University, Nanjing 210009, China

[6] Department of Radiology, Children's Hospital of Nanjing Medical University, Nanjing 210008, People's Republic of China

[7] Department of Neurology, First Affiliated Hospital of Nanjing Medical University, Nanjing 210029, China

**Table 1** Demographics of our dataset

| Characteristic | Local Hospitals | OASIS | |
|---|---|---|---|
| | AD (70) | AD (28) | HC (98) |
| Gender (M/F) | 24/46 | 9/19 | 26/72 |
| Age | 76.34 ± 7.81 | 77.75 ± 6.99 | 75.91 ± 8.98 |
| Socioeconomic Status | 2.89 ± 1.16 | 2.87 ± 1.29 | 2.51 ± 1.09 |
| Education | 2.63 ± 1.42 | 2.57 ± 1.31 | 3.26 ± 1.31 |
| MMSE | 21.12 ± 4.62 | 21.67 ± 3.75 | 28.95 ± 1.20 |
| CDR | 1 | 1 | 0 |

using biogeography-based optimization (BBO) method. Sun [9] proposed a multi-variate approach (MVA) to classify AD. Gorji and Haddadnia [10] employed pseudo Zernike moment (PZM) and scaled conjugate gradient (SCG) method. Their results showed PZM with the order of 30 performed the best. Du [11] continued to use PZM as features and replaced SCG with linear regression classifier (LRC).

In the last several years, deep learning [12, 13] has won success in nearly all fields of computer vision and image processing [14–16]. Hence, this paper analyzes applying deep learning method in the classification of AD. The hypothesis was that deep learning can assist in AD identification, and it can acquire better results than traditional machine learning approaches, since AlphaGo using deep learning has beaten all the AI programs. Among deep learning techniques, the convolutional neural network (CNN) was the most stable technique. Therefore, a CNN was constructed to classify AD from the healthy control (HC). The different pooling techniques used in CNN were tested and compared. The main difference of this study from past work, is that this study analyzes a deep learning technique. The CNN can extract task-related features (AD features in this study) automatically, and does not need the operations of users. Also, the CNN is known for its high accuracy, and it is equal to or even better than human experts.

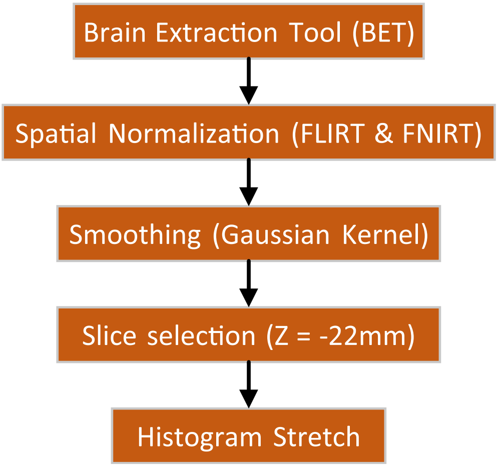

Brain Extraction Tool (BET)
Spatial Normalization (FLIRT & FNIRT)
Smoothing (Gaussian Kernel)
Slice selection (Z = -22mm)
Histogram Stretch

**Fig. 1** Pipeline of preprocessing

## Subjects and preprocessing

Two types of sources were used. One was from "Open Access Series of Imaging Studies" [17]. The dataset was made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN) [18, 19]. Subjects with missing records were removed. Finally, the subset with no missing records was selected, the dataset contained 28 AD patients and 98 HC subjects.

Since the OASIS dataset was imbalanced, 70 AD subjects from community advertisements were enrolled. Four local hospitals (Children's Hospital of Nanjing Medical University, Zhong-Da Hospital of Southeast University, Affiliated Nanjing Brain Hospital of Nanjing Medical University, and First Affiliated Hospital of Nanjing Medical University). This study was approved by the Ethics Committee of these local hospitals, and a signed informed consent form was obtained from every subject prior to entering this study.

The demographics of the dataset are listed in Table 1. In total, there was a balanced dataset of 98 AD patients and 98 HC subjects. This 196-image dataset is not a small dataset, since collecting AD patients and obtaining their permit is rather difficult. Grassi, Perna [20] used 123 mild cognitive impairment (MCI) subjects. Sheinerman, Toledo [21] collected 50 ADs, 50 frontotemporal dementia, 50 Parkinson's dementia, and 50 amyotrophic lateral sclerosis patients. In total, their dataset contains 200 subjects. Frolich, Peters [22] collected 115 MCI patients. All the above studies were carried out in the same period of this study.

For the 70 AD data provided by local hospitals, the scanning was implemented by a Siemens Verio Tim 3.0 T MR scanner (Siemens Medical Solutions, Erlangen, Germany). All subjects were to lie as still as possible with their eyes closed and were not to fall asleep. The imaging parameters were: TE = 2.48 ms, TR = 1900 ms, TI = 900 ms, FA = 9°, FOV = 256 mm × 256 mm, matrix size = 256 × 256, slice

**Table 2** Sizes of input, filter, and activation

| Operator | Width | Height | Channel |
|---|---|---|---|
| Input | *A* | *B* | *C* |
| Filter | *D* | *E* | *C* |
| Activation | *F* | *G* | |

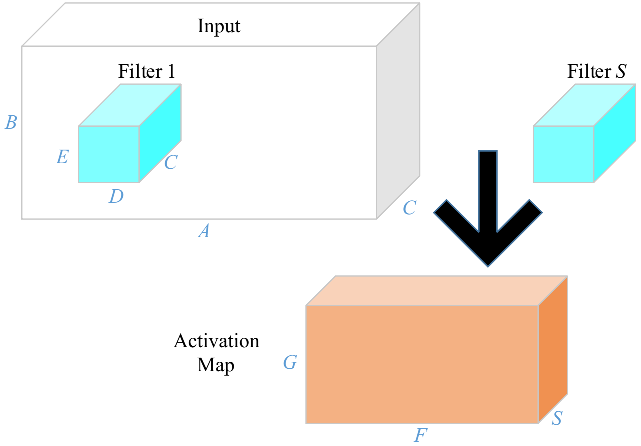


**Fig. 2** Illustration of convolution operation

thickness = 1 mm. Altogether 176 sagittal slices were acquired, via MP-RAGE sequence, in order to be coherent with OASIS dataset [9].

Following the same preprocessing procedure described in Sun [9], all the brain images were preprocessed through the pipeline as seenin Fig. 1. First, the brain extraction tool (BET) was employed to extract brain areas. FLIRT and FNIRT were used for spatial normalization. All the images were normalized to the MNI atlas. Smoothing was implemented by a Gaussian kernel. Slice selection was implemented at $Z = -22$ mm in MNI space, in order to include the hippocampus. Finally, histogram stretching (HS) [23] was used due to the two sources of brain images within our dataset. Since the data scanning was obtained at a different machine, the graylevel range may vary. Hence, HS is necessary to normalize the inter-scan images. The HS transformed original image $m$ to a new image $n$ as:

$$n(i,j) = \frac{m(i,j) - m_{\min}}{m_{\max} - m_{\min}} \tag{1}$$

where $m_{\min}$ and $m_{\max}$ represent the minimum (0%) and maximum (100%) intensity values of original image, respectively. Usually, 5% and 95% was used other than 0% and 100%, respectively. The reason was due to the pixels with least (0%) and greatest (100%) values are more vulnerable to noises, and choosing the 90% interval made the algorithm more reliable than using the 100% interval.

## Convolutional neural network

Traditional computer vision methods are composed of three important stages [24, 25]. First stage is feature 0extraction, second is feature reduction, and third is classification. Nevertheless, scholars combine these stages in standard convolutional neural network (CNN). That means, CNN does not need to set the feature manually. By contrary, the weights of its initial layers served as feature extraction, and their values were obtained by iterative learning. Also, CNN can obtain a better performance than peer classifiers, for example, feedforward neural network [26] and stochastic network [27]. The convolution layer performs feature extraction, and pooling layer performs reduction, and the softmax layer performs classification.

### Convolution layer

Convolution layer is the most important layer in the convolutional neural network (CNN) [28, 29]. It carries out a 2D convolution along scanner line for both 3D input and 3D filter. Suppose we have a 3D input with size of $A \times B \times C$, a 3D filter with size of $D \times E \times C$, then suppose the output activation map is with size of $F \times G$. For clearance, their sizes are listed in Table 2.

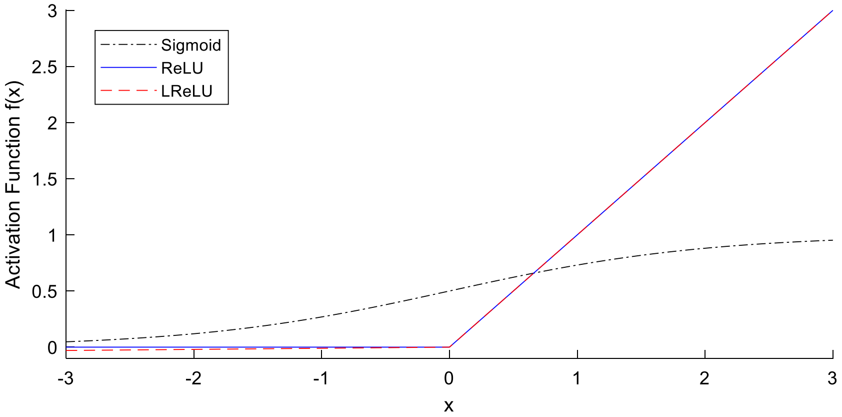
The value of $F$ and $G$ can be obtained by

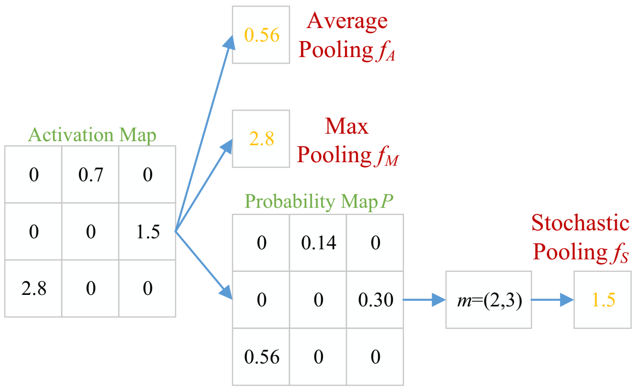
$$F = 1 + \frac{A - D + 2M}{Z} \tag{2}$$

$$G = 1 + \frac{B - E + 2M}{Z} \tag{3}$$

where $M$ represents margin, and $Z$ the stride size. Usually,there may be $S$ filters; hence, the activation map is $F \times G \times S$. An illustration is shown in Fig. 2.

**Fig. 3** Curves of three activation functions





| | | |
|---|---|---|
| 0 | 0.7 | 0 |
| 0 | 0 | 1.5 |
| 2.8 | 0 | 0 |

| | | |
|---|---|---|
| 0 | 0.14 | 0 |
| 0 | 0 | 0.30 |
| 0.56 | 0 | 0 |

**Fig. 4** A toy example of an activation map through average, max, and stochastic pooling

The purpose of the convolutional layer is in some degree equivalent to feature extraction in traditional pattern recognition task. The difference is the parameters are learnable and not fixed during training. The doc product results are called "*activation map*" or "*feature map*" in CNN jargon. The filters serve as feature detectors with weights being trainable.

## Activation function

Neurons in the activation map pass through a nonlinear activation function. Traditional activation function is in the form of sigmoid function [30], defined as

$$f_{sigmoid} = \frac{1}{1+\exp(-x)} \tag{4}$$

Recently, it was proven that the widespread saturation of sigmoidal function makes gradient-based learning and its variants perform poorly in the training neural network. To solve this problem, the rectified linear unit (ReLU) became popular [31], since it accelerated the convergence of stochastic gradient descent compared to the sigmoid function. Besides, it can be implemented by simply thresholding an activation map at zero. ReLU is defined as

$$f_{relu} = \max(0, x) \tag{5}$$

**Table 3** Division of training and test set

| | AD | HC | Total |
|---|---|---|---|
| Training | 49 | 49 | 98 |
| Test | 49 | 49 | 98 |
| Total | | | 196 |

**Table 4** Data augmentation

| | Total | AD | HC |
|---|---|---|---|
| Original Training Dataset | 98 | 49 | 49 |
| Rotation | 2940 | 1470 | 1470 |
| Gamma Correction | 2940 | 1470 | 1470 |
| Noise Injection | 2940 | 1470 | 1470 |
| Random Translation | 2940 | 1470 | 1470 |
| Scaling | 2940 | 1470 | 1470 |
| Random Affine | 2940 | 1470 | 1470 |
| Training Dataset after DA | 17,738 | 8869 | 8869 |

When their activation values are zero, the ReLU cannot learn via gradient-based methods, since the gradients are all zero. Therefore, a leaky ReLU (LReLU) [32] was proposed with definition of

$$f_{lrelu} = \begin{cases} x & x > 0 \\ 0.01x & \text{otherwise} \end{cases} \tag{6}$$

The differences of these three activation functions are drawn in Fig. 3.

## Pooling layer

The pooling layer served as the feature reduction stage in the traditional computer vision task. In addition, the pooling helped resist the effect of slight translation. In this study, the test three methods are average pooling ($f_A$), max pooling ($f_M$), and stochastic pooling ($f_S$). Suppose the pooling region is $Y$, the activation set $C$ included in $Y$ was

$$C = \{c_k | k \in Y\} \tag{7}$$

The average pooling [33] $f_A$ was defined as

$$f_A = \frac{\sum C_Y}{|C_Y|} \tag{8}$$

here $|x|$ is the cardinal number of set $x$. The max-pooling $f_M$ was defined as [34].

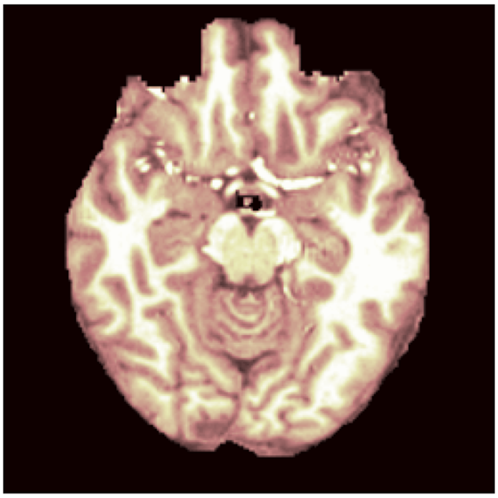
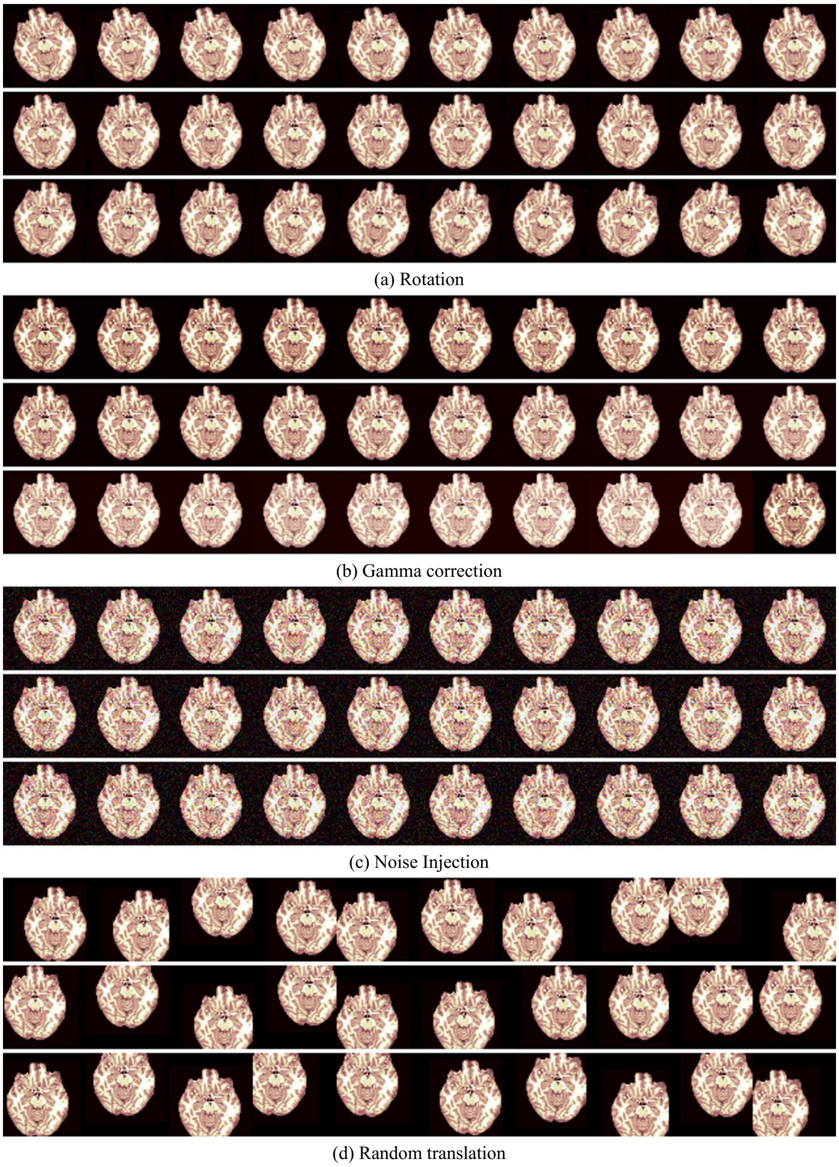


**Fig. 5** Original image (pink colormap was added for better vision performance)



(a) Rotation

(b) Gamma correction

(c) Noise Injection

(d) Random translation

**Fig. 6** Data augmented images

$$f_M = \max(C_Y) \tag{9}$$

Stochastic pooling selected the pooled map response [35] by sampling from a probability map $P = (p_1, \dots p_k, \dots)$, which is obtained via original activation map $c_k$ as

$$p_k = \frac{c_k}{\sum_C c_k} \tag{10}$$

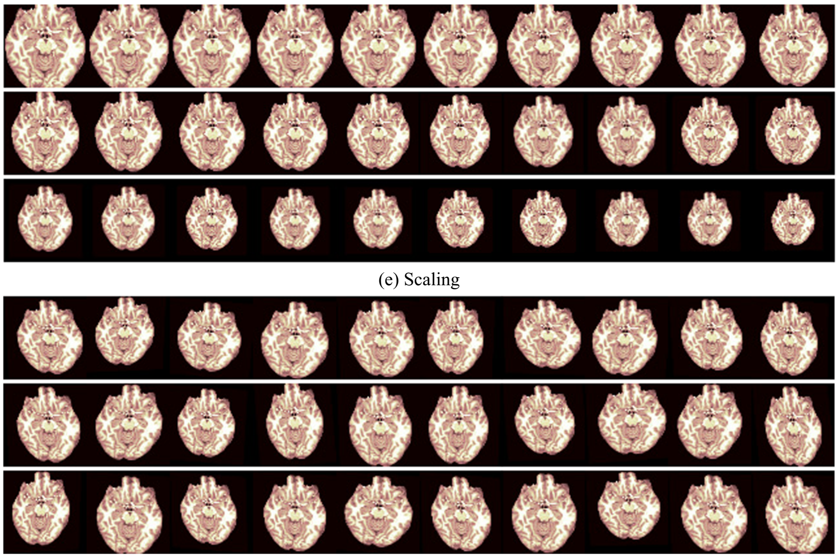
Then, the output of stochastic pooling $f_S$ was sampled from the multinomial distribution to pick a location $m$ within the activation region $C$ as

$$f_S = c_m, \text{where } m\sim(p_1, \dots, p_k, \dots) \tag{11}$$

A toy example of an activation map through three pooling methods is shown in Fig. 4. The average pooling $f_A$ produced a value of 0.56, which was the summation of the whole activation map divided by nine. The max pooling $f_M$ produced a value of 2.8, which was the maximum value of the activation map. For stochastic pooling $f_S$, probability map $P$ was first generated, and then randomly selected the position at (2,3). Finally, the output of $f_S$ was 1.5, which was the value at the second row and third column of the original activation map.

## Fully connected and softmax layer

The fully connected (FC) layer multiplied the input to the neurons within it by a fully connected weight matrix. Afterwards, a bias vector was added to the multiplication result. The fully-connected layers were usually followed by a softmax layer. Finally, the softmax layer applied the softmax function to the input.



(e) Scaling

(f) Random affine transform

**Fig. 6** (continued)

Suppose $Y(x|t)$ was the conditional probability of sample given class $t$, $T$ is the in-all number of classes, and $Y(t)$ was the class prior probability. The probability of a sample $x$ belonging to class $t$ can be deduced as:

$$Y(t|x) = \frac{Y(x|t) \times Y(t)}{\sum_{k=1}^{T} Y(x|k) \times Y(k)} \tag{12}$$

$N_t$ for simplicity was defined as

$$N_t = \ln(Y(x,t) \times Y(t)) \tag{13}$$

The simpler form of $Y(t|x)$ as

$$Y(t|x) = \frac{\exp(N_t(x))}{\sum_{k=1}^{T} \exp(N_k(x))} \tag{14}$$

## Experiment design

### Data augmentation

The 196-image dataset was not small to make a statistical analysis, but it was insufficient for a deep learning method. Th hold-out approach was used [36] to divide the whole dataset at random into the training and test sets as shown in Table 3. The training set contained 49 AD patients and 49 HC subjects. The test set contained the same number of AD and HC brains.

The 98-image training set was slightly small for a convolutional neural network to converge. Hence, the data augmentation (DA) method was used. The k-fold cross validation was not used and was not in cross validation because the whole dataset only contained 196 images, and this number

**Table 5** Feature extraction layers in proposed CNN structure

| Index | Layer Name | Filter/Pool Size | No. of Channels | No. of Filters | Stride | Padding |
|---|---|---|---|---|---|---|
| Image Input | | | | | | |
| 1 | Conv_ReLU_1 | 3 | 1 | 32 | 3 | 2 |
| 2 | Conv_ReLU_2 | 3 | 32 | 64 | 3 | 2 |
| | Pool | 3 | | | 1 | 1 |
| 3 | Conv_ReLU_3 | 3 | 64 | 128 | 3 | 0 |
| | Pool | 3 | | | 1 | 1 |
| 4 | Conv_ReLU_4 | 1 | 128 | 256 | 1 | 0 |
| | Pool | 3 | | | 1 | 1 |
| 5 | Conv_ReLU_5 | 1 | 256 | 512 | 1 | 0 |
| | Pool | 3 | | | 1 | 1 |
| 6 | Conv_ReLU_6 | 1 | 512 | 1024 | 1 | 0 |
| | Pool | 3 | | | 3 | 1 |

**Table 6** Classification layers in proposed CNN structure

| Index | Layer Name | Weights | Bias |
|---|---|---|---|
| 7 | FC_1 | 100 × 9216 | 100 × 1 |
| 8 | FC_2 | 2 × 100 | 2 × 1 |
| | Softmax | | |

was relatively small for a deep learning technique. The hold-out method could be used [36] in the 98-image training set, and the data augmentation method was used to enhance it to tens of thousands of images.

The first DA method was image rotation. The rotation angle $\theta$ was set from −15° to 15° in step of 1°. Thus, 30 new samples was created. The second DA method was gamma correction. The gamma-value $r$ varied from 0.7 to 1.3 with step of 0.02, again leading to 30 new samples. The third DA method was noise injection. 30 new noise-contaminated images were created for each original image. The zero-mean Gaussian noise with variance of 0.01 was employed. The fourth DA method used random translation, and 30 new randomly translated images were generated for each original image. The fifth DA method was scaling. The scaling factor $s$ varied from 0.7 to 1.3 with step of 0.02, providing 30 new images. The final DA method was random affine transform, in which 30 new randomly affined images were created for every original image. In total, 180 new images were made for each original image. The training dataset after DA contained 17,738 images, and half of which were AD patient and the rest were HC subjects, as shown in Table 4.

## Evaluation

The evaluation was reported on the 98-image test set. The ideal confusion matrix (CM) was

$$\mathrm{CM}_{ideal} = \begin{bmatrix} 49 & 0 \\ 0 & 49 \end{bmatrix} \tag{15}$$

**Table 7** Comparison of different activation and pooling methods

| Activation | Pooling | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| Sigmoid | Average | 93.06 ± 1.43 | 92.04 ± 1.16 | 92.55 ± 0.69 |
| Sigmoid | Max | 91.84 ± 1.92 | 91.02 ± 1.43 | 91.43 ± 1.20 |
| Sigmoid | Stochastic | 91.84 ± 1.92 | 92.86 ± 1.44 | 92.35 ± 1.54 |
| ReLU | Average | 95.71 ± 2.96 | 96.33 ± 2.69 | 96.02 ± 1.95 |
| ReLU | Max | 94.29 ± 2.32 | 97.14 ± 2.40 | 95.71 ± 1.72 |
| ReLU | Stochastic | 95.51 ± 3.57 | 96.33 ± 3.01 | 95.92 ± 2.45 |
| LReLU | Average | 95.71 ± 2.44 | 97.14 ± 2.40 | 96.43 ± 1.10 |
| LReLU | Max | 97.96 ± 1.36 | 97.35 ± 1.68 | 97.65 ± 0.97 |
| LReLU | Stochastic | 97.14 ± 2.19 | 96.94 ± 2.41 | 97.04 ± 1.40 |

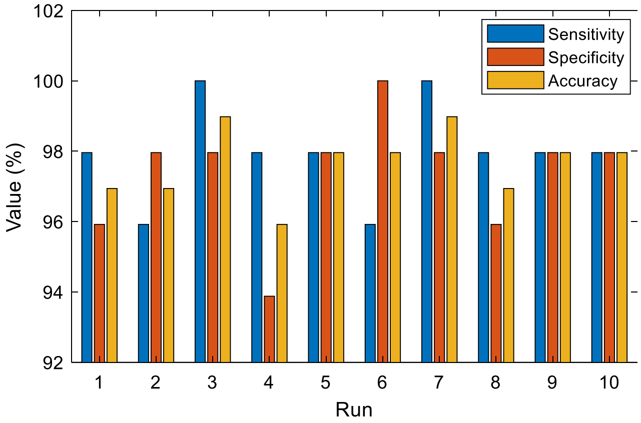


**Fig. 7** Performances of ten runs of LReLU-MP method

Suppose the confusion matrix was

$$\mathrm{CM}_{realistic} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} \tag{16}$$

where the positive class was AD patient and the negative class was healthy controls. The classification performance was evaluated by three commonly-used indicators: sensitivity, specificity, and accuracy. It was defined as

$$SEN = TP/(TP+FN) \tag{17}$$

$$SPC = TN/(TN+FP) \tag{18}$$

$$Acc = (TP+TN)/(TP+TN+FP+FN) \tag{19}$$

To avoid randomness, 10 runs were implemented, and the mean and standard deviation of the above three indicators were finally reported.

## Results

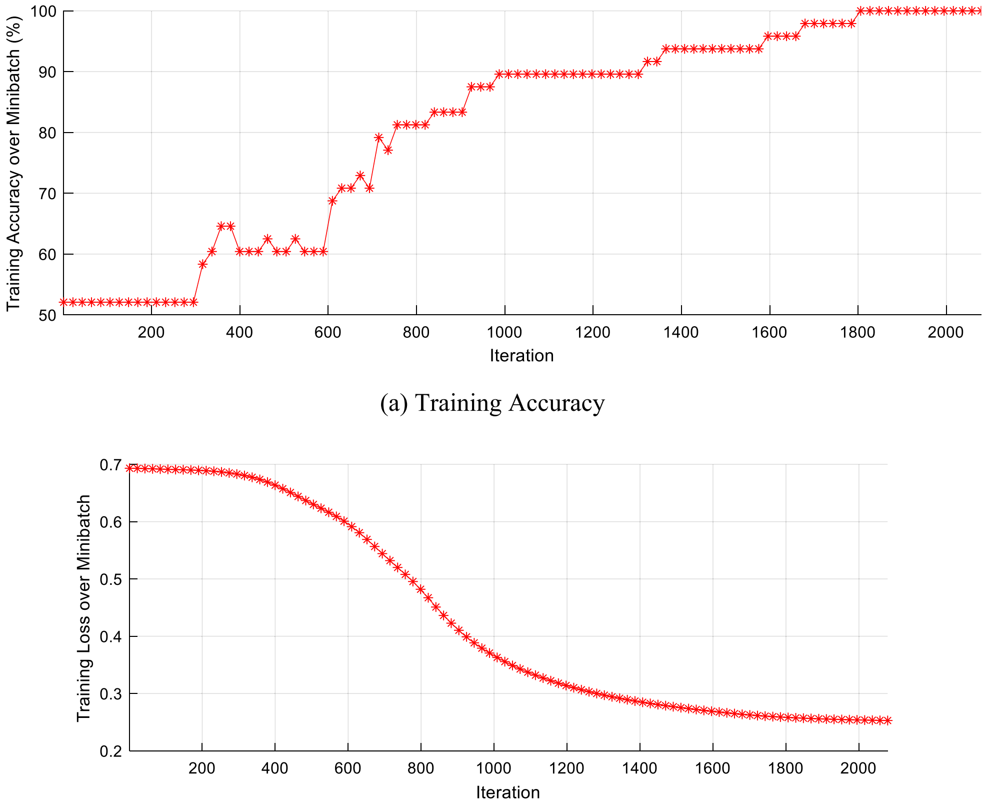
### Augmentation result

The image in Fig. 5 is an example, and shows the DA results of six different types. The augmented images after DA are shown in Fig. 6. Note that there were 30 new images for each DA type.

### Eight-layer CNN structure

The structure of the proposed CNN was adjusted by experience, and was tested as can be seen in "Optimal Structure" section. Finally, the 8-layer CNN structure, had only count those layers with learnable weights. Roughly speaking, six layers were contained in convolutional layers, serving as feature extraction, as listed in Table 5. Two fully connected layers are serving as classification, and their attributes were listed in Table 6.

**Fig. 8** Training performance within one run



(a) Training Accuracy

(b) Training Loss

## Test of activation and pooling methods

Different activation functions and pooling techniques were combined. The training was based on NVIDIA GeForce GTX 1050 with clock rate of 1455 MHz, compute capability of 6.1, and five multiprocessors. The training algorithm was the stochastic gradient descent with momentum (SGDM) [37].

The initial learning rate was set to 0.01, and would decrease by factor of 10 every 10 epochs. The maximum epochs was assigned with a value of 30. The momentum was set to 0.9. The size of minibatch was 256. The results of the different combination methods are listed in Table 7.

**Table 8** Comparison with other methods over the test set on average of 10 runs (Bold means the best, Unit: %)

| Algorithm | Dataset | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| BRC [4] | Private | 92.04 | 88.57 | 90.31 |
| TJM [5] | Private | 88.78 | 92.86 | 90.82 |
| RF [6] | Private | 87.76 | 90.82 | 89.29 |
| DF + SVM [7] | This | 83.67 | 87.96 | 85.82 |
| BBO [8] | This | 91.63 | 91.84 | 91.73 |
| MVA [9] | This | 92.65 | 92.86 | 92.76 |
| PZM + SCG [10] | This | 92.86 | 92.45 | 92.65 |
| PZM + LRC [11] | This | 93.06 | 92.86 | 92.96 |
| 8-layer LReLU-MP CNN (Proposed) | This | **97.96** | **97.35** | **97.65** |

## Our LReLU-MP method

Finally, the Leaky RELU and max-pooling (LReLU-MP) were selected for the method, since it got the highest performances among all nine different methods. The performances of its ten runs are shown in Fig. 7.

One run as an example, the curves of the training accuracy and training loss during the minibatch training are seen below in Fig. 8.

## Classification comparison

To further demonstrate the effectiveness of this proposed system of 8-layer convolutional neural network, it was compared with 8 existing algorithms, BRC [4], TJM [5], RF [6], DF + SVM [7], BBO [8], MVA [9], PZM + SCG [10], and PZM + LRC [11]. All those algorithms were briefly introduced in "Background" section. Some algorithms were used their own dataset, and were given a label of "private". Other algorithms were tested on the same dataset as this study, and were given a label of "this". The comparison results were shown in Table 8 and Fig. 9.

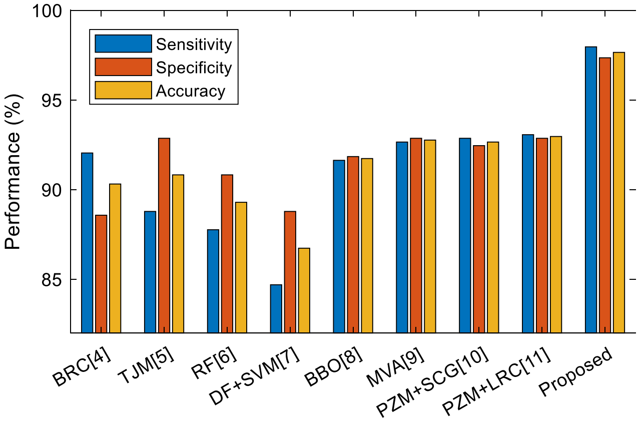


Fig. 9 Classifier Comparison

(BRC, brain region cluster; TJM, trace of Jacobian matrix; RF, random forest; DF, displacement field; SVM, support vector machine; BBO, biogeography-based optimization; MVA, multi-variate approach; PZM, pseudo-Zernike moment; SCG, scaled conjugate gradient; LRC, linear regression classifier; LReLU, leaky rectified linear unit; MP, max pooling; CNN, convolutional neural network)

## Optimal structure

The structure of the proposed CNN was obtained by experience (See "Eight-layer CNN Structure" section). Here we would like to give a quantitative experiment to show why six Conv_ReLU layers and two FC layers are optimal for this task. The first change is in the layer number of Conv_ReLU, note that a convolution layer is usually followed by a ReLU layer, and hence we combine them together. The average results of 10 runs over the test set are shown in Fig. 10.

Next, the six Conv_ReLU layers were kept, and tested the performances of using 1 FC layer, 2 FC layers, 3 FC layers, 4 FC layers, and 5 FC layers, respectively. The average results in terms of sensitivity, specificity, and accuracy are presented in Fig. 11.

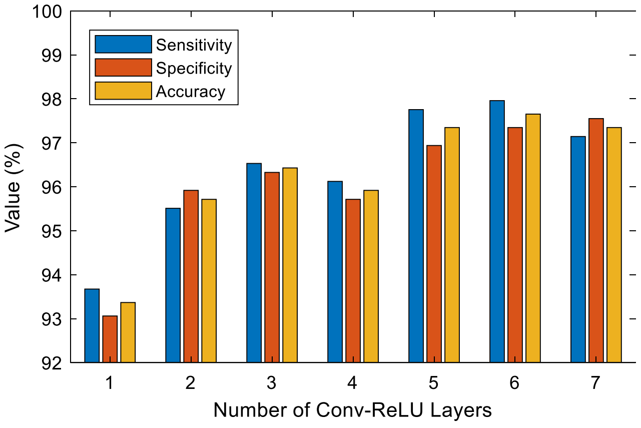


Fig. 10 Optimal number of Conv_ReLU layers

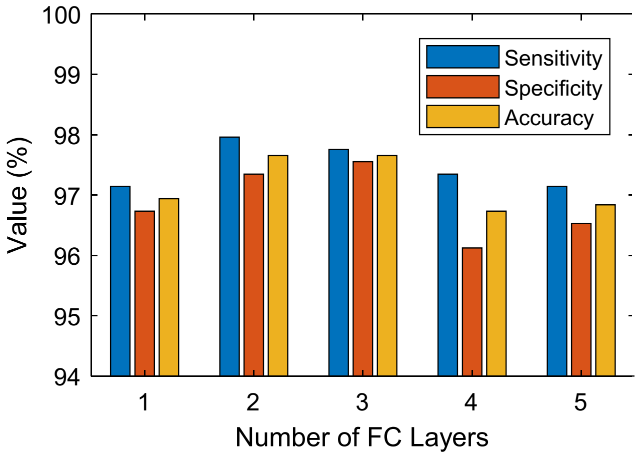


Fig. 11 Optimal number of FC layers

## Discussions

Table 7 showed that the best method was LReLU with max pooling method, which achieved a sensitivity of 97.96 ± 1.36%, a specificity of 97.35 ± 1.68%, and an accuracy of 97.65 ± 0.97%. The sigmoid activation related methods got the worst results, because the sigmoid function may become saturated when the absolute value of inputs is too large. The leaky ReLU gave the best result, because of two reasons: (i) The ReLU can accelerate the convergence of CNN; and (ii) the leaky mechanism makes the weights of neurons learnable even when the input values are less than zero.

In this study, the stochastic gradient descent with momentum (SGDM) was used, since it is the most common and reliable training method in deep learning. Nevertheless, there are other excellent training algorithms, such as Nesterov Momentum, AdaGrad [38], Newton's method, Adam, etc. In the future, the performances of those algorithms will be tested. Other technique like batch normalization will also be tested.

As observed in Fig. 8, the training accuracy increased gradually until perfect classification of 100%. There were several locations where the curve drops, due to the random selected 256 data each update time. On the other hand, the training loss decreased smoothly. Figure 8 also demonstrated that our selected hyperparameters were efficient, and could guarantee the CNN converge.

The results in Table 8 showed that this proposed 8-layer method achieved the highest sensitivity, specificity, and accuracy among all nine algorithms. For the sake of clear view, the results are pictured in Fig. 9. Here it can be observed that the proposed 8-layer LReLU-MP CNN increases the classification performance by at least 5% compared to traditional computer vision techniques. This demonstrates the superiority of the convolutional neural network, which is a typical tool of deep learning.

The curve in Fig. 10 shows as the performance increases as the number of Conv_ReLU layer increases. Nevertheless, the performance drops when there are 4 Conv_ReLU layer. When the layer number reaches to 5, the performance plateaus. We observe that the performance reaches the peak point when there are 6 Conv_ReLU layers.

FC layer is another important factor that influences the performance. Figure 11 shows that the accuracies with number of FC layers of 2 and 3 are equivalent. The number of 2 was chosen, since less layers cost less computation burden. Only 1 FC layer reduces the classification performance, and hence it was not used.

The clinical significance of our method is it can assist neuro-radiologists in giving an initial decision on the images. Presently, the computer vision is competent of or even better than human experts in many medical conditions, such as retinal disease [39], Parkinson's disease [36], etc. This paper gives an application of CNN in Alzheimer's disease. In the future, if portable MRI devices are available to the family, everybody can get MR images easily, and then diagnose with their cellphone apps using the proposed algorithm.

## Conclusions

This study proposed a novel deep-learning based approach to detect Alzheimer's disease. The experiments showed the effectiveness of the proposed method, and proved the hypothesis is true. Nevertheless, there remained several problems to be sorted with. First, the slice selection needs to be verified by more strict experiments. Second, the transfer learning needs to be considered to be used, because it can handle a small-size dataset more efficiently. Third, the hyperparameters were obtained by experience. In the future, we may test random search method to optimize the hyperparameters.

**Acknowledgements** This paper was supported by Open fund of Key Laboratory of Guangxi High Schools Complex System and Computational Intelligence (2016CSCI01), Natural Science Foundation of China (61602250) and Natural Science Foundation of Jiangsu Province (BK20150983), and National Institutes of Health (P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584).

## Compliance with ethical standards

**Conflict of interest** We have no conflicts of interest to disclose with regard to the subject matter of this paper.

## References

1. Lange, C. et al., Prediction of Alzheimer's Dementia in Patients with Amnestic Mild Cognitive Impairment in Clinical Routine: Incremental Value of Biomarkers of Neurodegeneration and Brain Amyloidosis Added Stepwise to Cognitive Status. *J. Alzheimers Dis.* 61(1):373–388, 2018.
2. Silveira, M. B. et al., F-18-Fluorocholine Uptake and Positron Emission Tomography Imaging in Rat Peritoneal Endometriosis. *Reprod. Sci.* 25(1):19–25, 2018.
3. Liu, G., Phillips, P., and Yuan, T.-F., Detection of Alzheimer's Disease by Three-Dimensional Displacement Field Estimation in Structural Magnetic Resonance Imaging. *J. Alzheimers Dis.* 50(1): 233–248, 2016.
4. Plant, C. et al., Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *NeuroImage.* 50(1):162–174, 2010.
5. Savio, A., and Grana, M., Deformation based feature selection for Computer Aided Diagnosis of Alzheimer's Disease. *Expert Syst. Appl.* 40(5):1619–1628, 2013.
6. Gray, K. R. et al., Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage.* 65:167–175, 2013.
7. Zhang, Y., Detection of Alzheimer's disease by displacement field and machine learning. *PeerJ.* 3:Article ID. e1251, 2015.
8. Wang, S.-H., Single slice based detection for Alzheimer's disease via wavelet entropy and multilayer perceptron trained by biogeography-based optimization. *Multimed. Tools Appl.*, 2016. https://doi.org/10.1007/s11042-016-4222-4.
9. Sun, J.-D., Multivariate Approach for Alzheimer's disease Detection Using Stationary Wavelet Entropy and Predator-Prey Particle Swarm Optimization. *J. Alzheimers Dis.*, 2017. https://doi.org/10.3233/JAD-170069.
10. Gorji, H. T., and Haddadnia, J., A novel method for early diagnosis of Alzheimer's disease based on pseudo Zernike moment from structural MRI. *Neuroscience.* 305:361–371, 2015.
11. Du, S., Alzheimer's Disease Detection by Pseudo Zernike Moment and Linear Regression Classification. *CNS Neurol. Disord. Drug Targets.* 16(1):11–15, 2017.
12. Raza, M. et al., Appearance based pedestrians' head pose and body orientation estimation using deep learning. *Neurocomputing.* 272: 647–659, 2018.
13. Bach-Andersen, M., Romer-Odgaard, B., and Winther, O., Deep learning for automated drivetrain fault detection. *Wind Energy.* 21(1):29–41, 2018.
14. Wei, G., Color Image Enhancement based on HVS and PCNN. *SCIENCE CHINA Inf. Sci.* 53(10):1963–1976, 2010.
15. Wu, L. N., Segment-based coding of color images. *Sci. China Ser. F-Inf. Sci.* 52(6):914–925, 2009.
16. Wu, L. N., Improved image filter based on SPCNN. *Sci. China Ser. F-Inf. Sci.* 51(12):2115–2125, 2008.
17. Ardekani, B. A., Figarsky, K., and Sidtis, J. J., Sexual Dimorphism in the Human Corpus Callosum: An MRI Study Using the OASIS Brain Database. *Cereb. Cortex.* 23(10):2514–2520, 2013.
18. Marcus, D. S. et al., Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19(9):1498–1507, 2007.
19. Marcus, D. S. et al., Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* 22(12):2677–2684, 2010.
20. Grassi, M. et al., A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. *J. Alzheimers Dis.* 61(4):1555–1572, 2018.
21. Sheinerman, K. S. et al., Circulating brain-enriched microRNAs as novel biomarkers for detection and differentiation of neurodegenerative diseases. *Alzheimers Res. Ther.* 9:13: Article ID. 89, 2017.
22. Frolich, L. et al., Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia. *Alzheimers Res. Ther.* 9:15: Article ID. 84, 2017.
23. Dhal, K. G., Quraishi, M. I., and Das, S., An Improved Cuckoo Search based Optimal Ranged Brightness Preserved Histogram

Equalization and Contrast Stretching Method. *Int. J. Swarm Intell. Res.* 8(1):1–29, 2017.

24. Lu, H. M., Facial Emotion Recognition Based on Biorthogonal Wavelet Entropy, Fuzzy Support Vector Machine, and Stratified Cross Validation. *IEEE Access.* 4:8375–8385, 2016.
25. Gorriz, J. M., and Ramírez, J., Wavelet entropy and directed acyclic graph support vector machine for detection of patients with unilateral hearing loss in MRI scanning. *Front. Comput. Neurosci.* 10: Article ID. 160, 2016.
26. Wu, L., Weights optimization of neural network via improved BCO approach. *Prog. Electromagn. Res.* 83:185–198, 2008.
27. Jun, Y., and Wei, G., Find multi-objective paths in stochastic networks via chaotic immune PSO. *Expert Syst. Appl.* 37(3):1911–1919, 2010.
28. Ozer, I., Ozer, Z., and Findik, O., Noise robust sound event classification with convolutional neural network. *Neurocomputing.* 272: 505–512, 2018.
29. Trakoolwilaiwan, T. et al., Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain-computer interface: three-class classification of rest, right-, and left-hand motor execution. *Neurophotonics.* 5(1):Article ID. 011008, 2018.
30. Chen, Y., Voxelwise detection of cerebral microbleed in CADASIL patients by leaky rectified linear unit and early stopping: A class-imbalanced susceptibility-weighted imaging data study. Multimed Tools Appl (2017). https://doi.org/10.1007/s11042-017-4383-9
31. Hara, K., Saito, D., and Shouno, H., *Analysis of Function of Rectified Linear Unit Used in Deep learning. in International Joint Conference on Neural Networks*. *IEEE*: Killarney, IRELAND. 144–151, 2015
32. Liew, S. S., Khalil-Hani, M., and Bakhteri, R., Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. *Neurocomputing.* 216:718–734, 2016.
33. Jiang, Y. et al., Cerebral Micro-Bleed Detection Based on the Convolution Neural Network With Rank Based Average Pooling. *IEEE Access.* 5:16576–16583, 2017.
34. Hang, S. T., and Aono, M., Bi-linearly weighted fractional max pooling An extension to conventional max pooling for deep convolutional neural network. *Multimed. Tools Appl.* 76(21): 22095–22117, 2017.
35. Lv, Y.-D., and Sui, Y., Alcoholism detection by data augmentation and convolutional neural network with stochastic pooling. *J. Med. Syst.* 42(1):Article ID. 2, 2018.
36. Camps, J. et al., Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. *Knowl.-Based Syst.* 139:119–131, 2018.
37. Wawrzynski, P., ASD plus M: Automatic parameter tuning in stochastic optimization and on-line learning. *Neural Netw.* 96:1–10, 2017.
38. Hadgu, A.T., Nigam, A., and Diaz-Aviles, E., *Large-Scale Learning with AdaGrad on Spark. in International Conference on Big Data*. *IEEE*: Santa Clara, CA. 2828–2830, 2015
39. Al-Bander, B. et al., Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. *Biomed. Signal Process. Control.* 40:91–101, 2018.