Paper:

Image Classification of MRI for Alzheimer's Disease Using Two Deep Neural Networks and their Explainability

Boyue Song, Shinichi Yoshida

*Graduate School of Engineering, Kochi University of Technology, Kami, Kochi 782-8502, Japan
E-mail: 258005p@gs.kochi-tech.ac.jp
**School of Information, Kochi University of Technology, Kami, Kochi 782-8502, Japan
E-mail: yoshida.shinichi@kochi-tech.ac.jp

Currently, many studies have realized the automatic classification of various neurological diseases with a synergy of raw fMRI data and convolutional neural networks (CNNs). However, CNN model has been criticized for its uninterpretability as a black box, which can greatly affect the credibility of the model, especially in the field of clinical diagnosis. In this paper, we used two modified CNNs to classify different stages of Alzheimer's disease (AD) and applied Gradients-weighted Class Activation Mapping (Grad-CAM) on these models to explain them. However, the images of Grad-CAM didn't highlight different regions of the brains according to the stages of AD as expected. Instead, the highlighted regions appeared randomly, which indicated the models didn't effectively extract useful higher-level features for the different stages of AD. Instead, specific lower-level features of the patients' brain might become the basis for classification.

Keywords: Alzheimer's disease; convolutional neural networks; clinical functional magnetic resonance imaging; Gradients-weighted Class Activation Mapping; visual explanation.

1. Introduction

Since the early 1990s, functional magnetic resonance imaging (fMRI) has become a common technique in brain mapping research because it can provide rich functional information without injections, surgery or exposure to ionizing radiation. fMRI collects functional information by detecting the blood-oxygen-level dependent (BOLD) contrast, discovered by Seiji Ogawa in 1990 [1]. This is a specialized scan of the brain and the body. It maps neural activity in the brain or spinal cord of humans and other animals by imaging the change in blood flow (hemodynamic response) response to energy use by brain cells [2].

Although fMRI affords such strong advantages, it is not yet widely used for clinical diagnosis due to its high susceptibility to noise and the extremely high complexity of its data. Typically, the useful signal variation is only 2% to 5% of the signal strength. Hence, even a small amount of noise can have a big impact on the data. Besides, a set of fMRI data can contain millions of data points because of the voxel-by-voxel scanning in both space and time. Therefore, in scientific research, statistical methods are usually used to analyze fMRI data, such as independent components analysis (ICA) [3] and General Linear Model (GLM) [4] [5].

However, handcrafted feature extraction is necessary for traditional statistical methods, which is often accompanied by inefficiencies and occasional mistakes. Recently, improvements in the processing power of computer hardware and innovation of the latest graphical software, have led researchers to focus increasingly on automation that is informed by deep-learning.

Recently, many studies have realized the automatic classification of various neurological diseases with a synergy of raw fMRI data and convolutional neural networks (CNNs). For example, Sarraf et al. [6] successfully classified fMRI data of Alzheimer's subjects from normal controls with the accuracy up to 96.85%. These scans were pre-processed by standard pipeline. After that, 4D fMRI data were converted to 2D images across z and t axes and divided into training and testing dataset randomly. Finally, LeNet-5 was used for classification. Similarly, Ramzan et al. [7] used almost the same preprocessing pipeline but replaced the network with residual neural networks to perform AD classification task. The accuracy was as high as 97.88% with fine-tuned transfer learning. In another study, Bi et al. [8] designed two networks for AD classification. Convolutional learning method and recurrent learning method were applied for deep regional-connectivity features and deep adjacent positional features respectively. After that, the extreme learning machine (ELM) was added to both networks to improve the learning ability. Parmar et al. [9] applied a modified 3D CNN to resting-state fMRI for classification of AD. In this method, 4D fMRI data is sliced along the time axis to facilitate the training. The accuracy was almost up to 98.96%. Naz et al. [10] employed architectures of convolutional networks using freeze features extracted from source data set ImageNet (AlexNet, GoogleNet, ResNet, DenseNet, Inceptionv3, Inception-ResNet) for binary and ternary classification to classify ADNI dataset, and achieves an indentification test accuracy of 99.27% (MCI/AD), 98.89% (AD/CN) and 97.06%

(MCI/CN).

However, deep learning models are often mentioned as black boxes due to their uninterpretability. This means that we can't know the reasons why any given prediction is made. This is true regardless of whether the prediction is correct or not, which can greatly affect the credibility of the model, especially in the field of clinical diagnosis. In 2016, Zhou et al. [11] proposed a method named Class Activation Mapping (CAM) to visualize the model with the help of feature maps of the last convolutional layer. This method enables visualization of the basis for the results of the model classification, including both correct and incorrect cases. In 2017, Ramprasaath et al. [12] proposed a new method named Gradient-weighted Class Activation Mapping (Grad-CAM) on the above method by combining gradients of the Gradient Descent Algorithm and feature maps of the last convolutional layer. As shown in Fig. 1, Grad-CAM can localize class-discriminative regions by which the network makes predictions.

AD is a kind of neurological diseases affected by memory-related neurodegeneration. The brain regions of different stages of AD show different patterns of blood oxygenation levels [13] [14]. These information can be recorded by fMRI. Since the models classified different stages of AD clearly, the Grad-CAM heat-map should localize different certain regions according to different stages of AD.

In this paper, we applied Grad-CAM on two modified models which were used to classify different stages of AD. The first network was modified residual neural network based on Resnet-18 and the second one was a modified 3D-CNN. The images of Grad-CAM didn't highlight different regions of the brains according to the different stages of AD as expected. Instead, the highlighted regions appeared randomly, which indicated the networks didn't effectively extract useful features for the different stages of AD. Instead, specific features of the patients' brain might become the basis for classification.

2. Method

2.1. Dataset and preprocessing

All the resting-state fMRI data used for this study were acquired from the ADNI online database [15]. The dataset of Res-Net included 6 stages of AD named CN, SMC, EMCI, MCI, LMCI, AD and 10 subjects for each class. The dataset of 3D-CNN included 4 stages of AD named CN, EMCI, LMCI, AD and 30 subjects for each class. A standard pipeline of data preprocessing including various steps was applied on the dataset to remove noise and improve generality.

Firstly, the dataset was converted from DICOM format to NIFTI format with a toolbox named dcm2niix. After that, Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL) [16] [17] is applied for all other preprocessing steps. Brain extraction was used on both anatomical data and functional data with FSL-BET toolbox. Then, motion correction was applied to remove the effect of subjects' head motion during data acquisition with FSL-MCFLIRT. Slice timing correction was also performed to make temporal adjustment. Intensity normalization was performed to keep each volume at the same mean intensity. Spatial smoothing was applied to remove noise with a 5 mm FWHM Gaussian kernel. After that, temporal high-pass filtering was applied to remove low-frequency noise with a temporal filter of cut-off frequency 0.01 Hz. Finally, spatial normalization was performed to register the fMRI images from subjects' space to a standard space of MNI152 template.

After performing all the preprocessing steps on fMRI data, for each subject, a $91 \times 109 \times 91 \times 140$ fMRI data matrix was obtained in which each volume contained $91 \times 109 \times 91$ data points for 140 time points.

For Resnet-18, these 4D fMRI scans were sliced into 2D images along with image height and time axis. We took the middle 30 slices along the height axis as they contained the most functional information. Thus, 4200 images were obtained for each subject. Totally, there were 252000 images for all the subjects. We divided them randomly and took 70% of them as the training set, 20% as the validation set, and 10% as the test set. Besides, in order to keep fMRI images suitable for Resnet-18, each image was first cut to a size of 91×91 and then enlarged to a size of 224×224.

For 3D-CNN, we sliced these 4D fMRI scans into smaller 4D matrix along time axis. Each sample has 5 time points as a $91 \times 109 \times 91 \times 5$ 4D matrix. 20 samples were extracted from the fMRI data of each subject thus 2400 samples for all the subjects. After that, these 4D samples were divided randomly with 1600 samples for training set, 400 samples for validation set and 400 samples for test set. In order to keep 3D fMRI data suitable for 3D-CNN, each sample was cut to a size of $91 \times 91 \times 91$.

2.2. Grad-CAM

Grad-CAM [12] is an improved method based on CAM [11]. Gradient information flowing into the last convolutional layer of the CNN was used to understand the importance of each neuron for a decision of interest.

Firstly, the gradients of the score for class c, y^c, with respect to feature maps A^k of the last convolutional layer, *i.e.* $\frac{\partial y^c}{\partial A^k}$ are computed. After that, global average pooling function is applied to these gradients to obtain the neuron importance weights α_k^c:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad \cdots\cdots\cdots\cdots \quad (1)$$

The weight α_k^c means the importance of feature map k for a target class c. Z means the number of elements in A^k.

Then, a weighted combination of forward feature maps is performed. Finally, a ReLU is used to remove the neg-

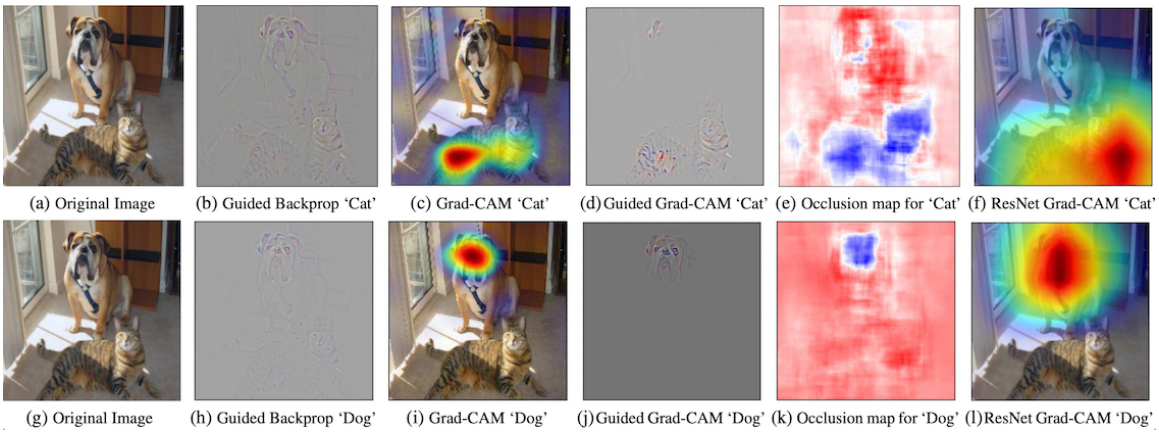


Fig. 1. Grad-CAM [12]

ative values:

$$L^c_{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \quad \ldots\ldots \quad (2)$$

A coarse heat-map which has the same size as the feature maps of the last convolutional layer is obtained. As mentioned in [12], the size of the last feature maps is usually 14×14 in some famous networks such as VGG [18] and AlexNet [19]. After we get the coarse Grad-CAM heat-map, bilinear interpolation algorithms are needed to ensure the same size between the heat-map and the original image so that we can get an intuitive comparison.

2.3. Modified networks

We added Grad-CAM to two networks which were used for classification of different stages of AD. As mentioned above, the coarse heat-map is enlarged to keep it and original image the same size. Originally, the size of the last forward feature maps is too small to get good results after interpolation. Thus, we make some changes on the original networks to enlarge the last feature maps without compromising accuracy.

For Resnet-18 used for AD multi-classification, the structure is described in detail in [7]. In our work, the stride of the 13th convolutional layer was changed to 1. In this case, the size of the last feature maps was kept as a size of 14×14 which is suitable for Grad-CAM. The batch normalization layer and the ReLU activation layer after the last convolutional layer were removed by which the process of gradients flowing back was simplified. Finally, we replaced the penultimate linear layer with a global average pooling layer which acts as a structural regularizer, preventing over-fitting during training [20].

For 3D-CNN, detailed information can be found in [9]. We made four changes on the original network. First, we added another 3D convolutional layer after the 6th convolutional layer. The kernel size was $2 \times 2 \times 2$ and the stride was $1 \times 1 \times 1$. Second, ReLU activation layer were added after every convolutional layer, not the certain layers. Third, the third and fourth max pooling layers were removed to keep the last feature maps as a size of $14 \times 14 \times 14$. Finally, the global average pooling layer was also added to replace the penultimate linear layer.

3. Results

The networks were built and trained using Python 3.6[1] with pytorch 1.4.0.[2] on a Linux machine with 256 GB RAM and 8 GB NVIDIA GPU card[3]. The training parameters were consistent with [7] and [9] respectively. As shown in Table. 1, the accuracy of both two CNNs was as high as 99% after training. Fig. 2 shows the graphical trend of loss and accuracy with 3D-CNN. The training loss and accuracy of ResNet-18 are not displayed since the model converges within a few epochs.

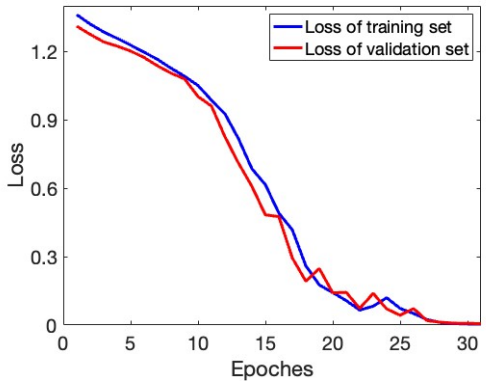
Fig. 3 shows the Grad-CAM heat-maps with respect to different classes of ResNet-18. From top to bottom are different patients belong to CN, SMC, EMCI, MCI, LMCI, AD. From left to right are Grad-CAM images belong to each class of the model. As we can see, only for the correct class, there is a large highlighted region which is localized by Grad-CAM. For other five classes, the highlighted region is smaller or not even. Therefore, the network makes correct predictions based on these large highlighted region.

Fig. 4 shows the Grad-CAM heat-map of patients corresponding to the ground truth. As we can see, the regions localized by Grad-CAM appears randomly. The whole brain is highlighted in some images, such as the second image in Fig. 4(a), the second image in Fig. 4(b) and the second image in Fig. 4(c). In some images, The front or back half of the brain is highlighted, such as the last image of Fig. 4(a), the fourth image of Fig. 4(c), the fourth image of Fig. 4(d) and the first image of Fig. 4(f). Besides, there is only a small region is localized in some images, such as the first image of Fig. 4(c), the second image of Fig. 4(f).

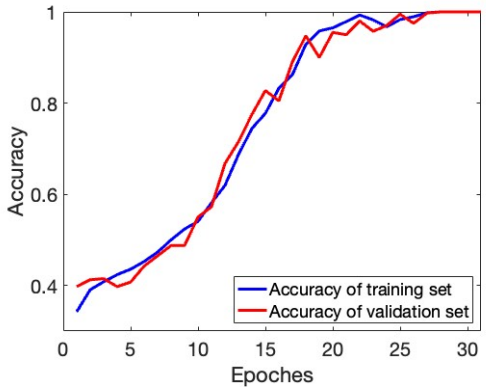
1. https://www.python.org/
2. https://pytorch.org/
3. https://www.nvidia.com/

Table 1. The accuracy of test set with two CNNs.

	CN	SMC	EMCI	MCI	LMCI	AD	Total
ResNet-18	0.9992	0.9997	0.9992	1.0000	0.9980	0.9990	0.9992
3D-CNN	0.9899	-	0.9806	1.0000	-	1.0000	0.9925



(a) Graphical trend of loss with 3D-CNN



(b) Graphical trend of accuracy with 3D-CNN

Fig. 2. Graphical trend of loss and accuracy with 3D-CNN.

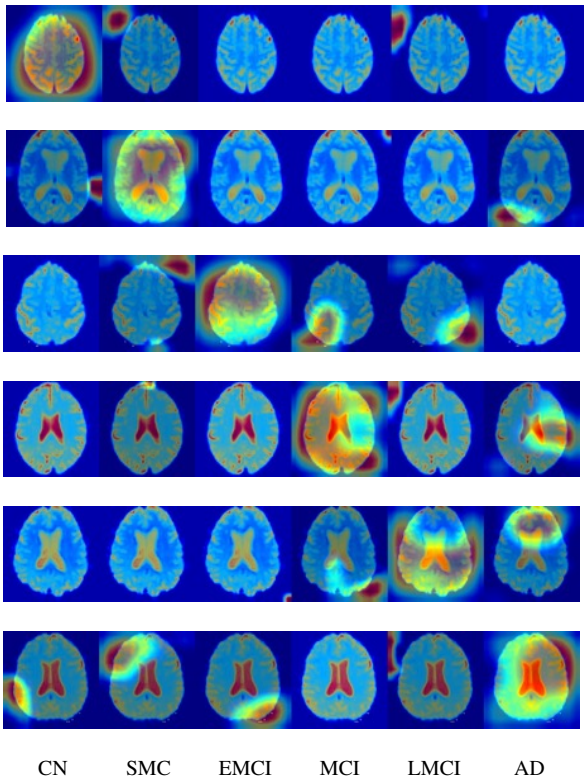


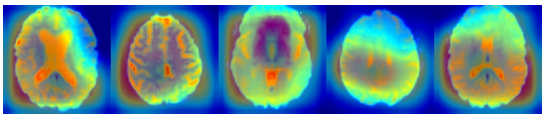
Fig. 3. Grad-CAM of patients corresponding to all classes. From top to bottom are six patients belong to CN, SMC, EMCI, MCI, LMCI, AD.

Although these fMRI images are classified accurately, the basis by which they are classified is inconsistent. That means the useful features of different stages of AD are not extracted effectively by the model. Instead, the model finishes the task based on other kinds of features, such as the shapes and the relative values of different pixels of the brains.

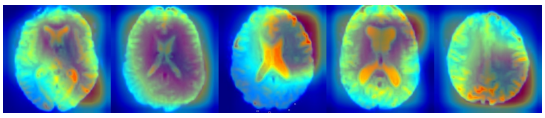
Fig. 5 shows several Grad-CAM heat-maps of incorrect prediction samples. For example, the brain in the first row images belongs to a CN patient but the network classified it to the LMCI class. As shown in the image, the localized region in the fourth image is larger than that in the first image, which explains why an incorrect result is obtained. The network predicts the brain in the second row images as a MCI patient, but it actually belongs to a EMCI patient. For the third row images and the fourth row images, they actually belong to LMCI and AD class respectively. However, the network classifies them both to the CN class. This can be seen clearly from the Grad-CAM heat-maps.

We also applied Grad-CAM to the 3D-CNN. The results are shown in the Fig. 6. As shown in the images, the localized regions are different even for the patients who belong to the same class.

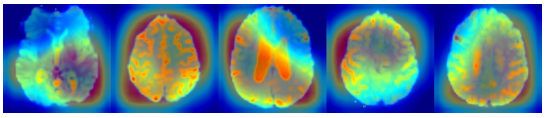
As shown in the Fig. 3, the differences between different classes are readily recognized since only the correct heat-map has a large highlighted region. As shown in the Fig. 4 and Fig. 6, the images of Grad-CAM don't localize different regions of the brains according to the stages of AD as expected. Instead, the highlighted regions appear randomly, which indicates that the networks classifies the dataset not the features of different stages of AD, but some other lower-level features of certain subjects, such as the shape of the brains. It is supposed that this happens due to the inappropriate division of the dataset. As mentioned above, both [9] and [7] sliced the fMRI data into smaller



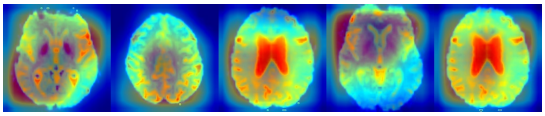
(a) Grad-CAM of CN



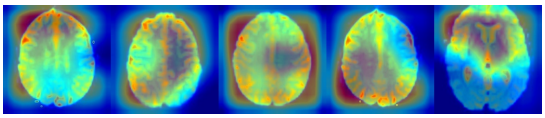
(b) Grad-CAM of SMC



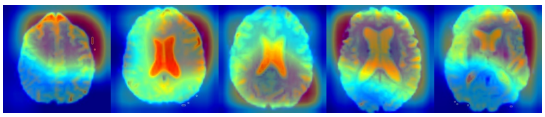
(c) Grad-CAM of EMCI



(d) Grad-CAM of MCI

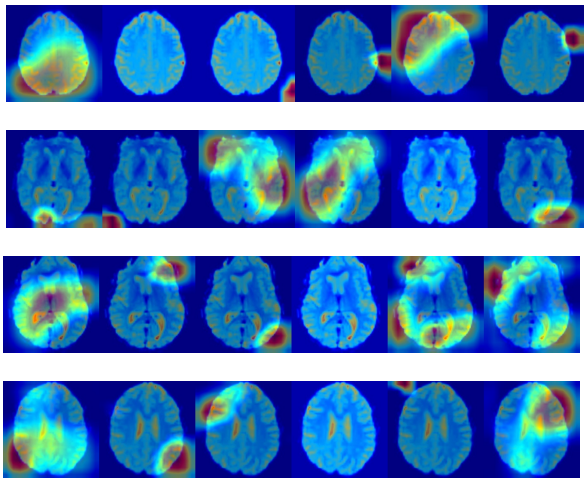


(e) Grad-CAM of LMCI



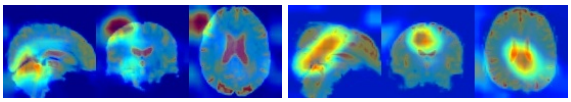
(f) Grad-CAM of AD

Fig. 4. Grad-CAM of patients corresponding to the ground truth. From top to bottom are 30 patients belong to CN, SMC, EMCI, MCI, LMCI, AD.

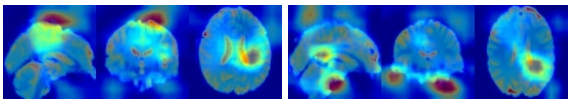


CN SMC EMCI MCI LMCI AD

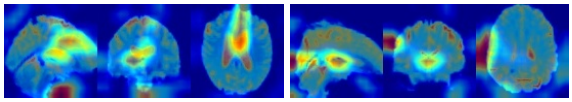
Fig. 5. Grad-CAM of incorrect predictions. From top to bottom are four patients belong to CN, EMCI, LMCI, AD.



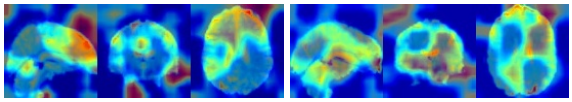
(a) CN



(b) EMCI



(c) LMCI



(d) AD

Fig. 6. Grad-CAM of 3D-CNN. From top to bottom are eight patients belong to CN, EMCI, LMCI, AD.

parts firstly. After that, these sliced data were mixed and then divided into training set, validation set and test set randomly, which caused a problem: for a certain subject, its fMRI data with different spatial or temporal coordinate were divided to the training set and test set. As shown in the Grad-CAM, the networks extracted lower features corresponding to the certain subjects, but not the higher-level features, like the information of the patterns of blood oxygenation levels of different stages of AD.

4. Conclusion

In this paper, we applied Grad-CAM to two networks for AD classification to get visual explanations. the Grad-CAM heat-maps of different classes were easily to recognized since only the correct image was localized with a large region. However, it is supposed that because of inappropriate division of the dataset, the samples belong to the same subjects were divided both into training set and test set. In this case, Grad-CAM showed that the networks classified the samples based on some lower level like the shape of the brains but not any higher-level features.

For further research, the inappropriate division of dataset should be avoided. Besides, as far as we know, it is hard for the CNNs to classify diseases based on raw fMRI data. Thus, a synergy of feature extraction methods and CNNs is necessary.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP22K12786, JP21H03553, JP20H00267 and the Chinese Scholarship Council.

References:

[1] S. Ogawa, T. Lee, A. S. Nayak, and P. Glynn, "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields", Magnetic Resonance in Medicine, 14, 1990.

[2] S. Huettel, A. Song, and G. McCarthy, "Functional Magnetic Resonance Imaging, Second Edition", 12 2008.

[3] M. A. Oghabian, S. A. H. Batouli, M. Norouzian, M. Ziaei, and H. Sikaroodi, "Using functional magnetic resonance imaging to differentiate between healthy aging subjects, Mild Cognitive Impairment, and Alzheimer's patients", Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences, 15(2):84, 2010.

[4] M. Monti, "Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach", Frontiers in Human Neuroscience, 5, 2011.

[5] C. R. Pernet, "The General Linear Model: Theory and Practicalities in Brain Morphometric Analyses", pp. 75–85, 2018.

[6] S. Sarraf and G. Tofighi, "Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks", arXiv preprint arXiv:1603.08631, 2016.

[7] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks", Journal of medical systems, 44(2):1–16, 2020.

[8] X. Bi, X. Zhao, H. Huang, D. Chen, and Y. Ma, "Functional brain network classification for Alzheimer's disease detection with deep features and extreme learning machine", Cognitive Computation, 12(3):513–527, 2020.

[9] H. Parmar, B. Nutter, R. Long, S. Antani, and S. Mitra, "Spatiotemporal feature extraction and classification of Alzheimer's disease using deep learning 3D-CNN for fMRI data", Journal of Medical Imaging, 7(5):056001, 2020.

[10] S. Naz, A. Ashraf, and A. Zaib, "Transfer learning using freeze features for Alzheimer neurological disorder detection using ADNI dataset", Multimedia Systems, 28(1):85–94, 2022.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929, 2016.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", In Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[13] M. J. Prince, A. Wimo, M. M. Guerchet, G. C. Ali, Y.-T. Wu, and M. Prina, "World Alzheimer Report 2015-The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends", 2015.

[14] A. Association et al., "2018 Alzheimer's disease facts and figures", Alzheimer's & Dementia, 14(3):367–429, 2018.

[15] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods", Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 27(4):685–691, 2008.

[16] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl", Neuroimage, 62(2):782–790, 2012.

[17] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith, "Bayesian analysis of neuroimaging data in FSL", Neuroimage, 45(1):S173–S186, 2009.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, 25, 2012.

[20] M. Lin, Q. Chen, and S. Yan, "Network in network", arXiv preprint arXiv:1312.4400, 2013.