Differentiating Alzheimer's Disease from Cognitively Normal Individuals Using Convolutional Neural Networks: A Reproducible Study

Elina Thibeau-Sutre[1], Camille Brianceau[2], and Ninon Burgos[2]

[1] Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands
e.thibeau-sutre@utwente.nl
[2] Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
ninon.burgos@cnrs.fr

Abstract. Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects millions of individuals worldwide. Early and accurate diagnosis is crucial for effective intervention and patient care. This study aims to develop a reproducible deep learning model based on convolutional neural networks (CNNs) to differentiate between AD patients and cognitively normal participants using brain MRI scans.
The chosen CNN model demonstrated a high level of accuracy in distinguishing AD patients from cognitively normal controls. On the test set, the model achieved a balanced accuracy of 92.3%. The attribution maps associated to the trained network highlighted regions known to be affected by the disease (medial temporal regions).
This reproducible study demonstrates the potential of convolutional neural networks in effectively differentiating Alzheimer's disease patients from cognitively normal individuals based on brain MRI scans. The high balanced accuracy achieved by the model highlight its clinical relevance and potential as a valuable diagnostic tool. The open-source code used in this study is made available to facilitate further research and ensure transparency and reproducibility in the field of neuroimaging-based AD diagnosis.

Keywords: Alzheimer's disease · Deep Learning · Magnetic Resonance Imaging.

1 Introduction

Alzheimer's disease (AD) affects over 20 million people worldwide. Neuroimaging provides useful information to identify AD [1], such as the atrophy due to gray matter loss with anatomical magnetic resonance imaging (MRI). A major interest is then to analyze those markers to identify AD at an early stage. Machine learning and deep learning methods have the potential to assist in identifying

patients with AD by learning discriminative patterns from neuroimaging data [2].

As the most widely used architecture of deep learning, convolutional neural networks (CNN) have attracted huge attention thanks to their great success in image classification [3]. Contrary to conventional machine learning, deep learning allows the automatic abstraction of low-to-high level latent feature representations. Thus, one can hypothesize that deep learning depends less on image preprocessing and requires less prior on other complex procedures, such as feature selection, resulting in a more objective and less bias-prone process [4].

The purpose of this paper is to explain the results of a deep learning network trained to differentiate Alzheimer's disease patients from cognitively normal participants. The source code for the experiments and models described in this paper will be made available on GitHub and is attached to this submission during the review process.

2 Materials

Our classifier was trained and evaluated on the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set. More specifically, our population includes 2 labels:

CN Cognitively Normal images corresponds to the baseline image of participants who were always diagnosed as CN,

AD Alzheimer's Disease images corresponds to the baseline image of participants who were always diagnosed as AD.

The stability of the diagnosis was established by considering the first 3 years of follow-up only. Participants with less than 3 years of follow-up or no T1w-MRI with N3 preprocessing at baseline were excluded from the data set. Table 1 summarizes the characteristics of our different populations.

Table 1: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores at baseline for ADNI. Values are presented as mean (standard deviation) [range]. M: male, F: female

	Images	Age	Gender	MMSE	CDR
CN	330	74.4 (5.8) [59.8, 89.6]	160 M / 170 F	29.1 (1.1) [24, 30]	0: 330
AD	336	75.0 (7.8) [55.1, 90.9]	185 M / 151 F	23.2 (2.1) [18, 27]	0.5: 160; 1: 175; 2: 1

3 Methods

3.1 Preprocessing of T1-weighted MRI

Several scans are provided for each time point of each participant in ADNI. For each participant we chose to use its baseline image which was already preprocessed by the N3 algorithm by the data set provider. Then the N4ITK method

[5] was used for bias field correction. Next, a linear (affine) registration was performed using the SyN algorithm from ANTs [6] to register each image to the MNI space (ICBM 2009c nonlinear symmetric template) [7,8]. To improve the computational efficiency, the registered images were further cropped to remove the background. The final image size is 169×208×179 with 1 mm3 isotropic voxels. Finally intensities were rescaled to [0,1] based on min and max values.

3.2 Deep learning network

Our network takes as input the whole 3D image and outputs a value for each of the label (AD and CN) which can be seen as probabilities to belonging to each of these classes after applying a SoftMax.

Hyperparameter search We performed a random search to find the best possible network hyperparameters [9]. The following parameters were fixed:

- a convolutional block consists of:
  1. a convolutional layer with a kernel size of 3, padding of 1 and stride of 1. The first number of channels is 16, the number of channels in one of the following layers is twice the number of channels in the previous layer.
  2. (Optional) a normalisation layer
  3. an activation layer
  4. a max pooling layer of stride and kernel of size 2.
- the number of nodes of the final fully-connected (FC) layer is 2. The number of nodes of intermediate FC layers is computed so that the ratio between the input and output size is the same for every FC layer. An activation layer follows every FC layer except the last one.
- training hyperparameters were fixed as described in paragraph 3.2, except for the learning rate which can vary.

If a normalisation layer is chosen, it is the same in all convolutional blocks. Similarly the same activation function is used after each convolutional or FC layer. The following parameters can vary: the number of convolutional blocks (2 to 6), the normalisation layer (batch, instance or none), the activation layer (ReLU, leaky ReLU or SeLU), the number of FC layers (1 to 4), the learning rate (0.01 to 0.0001).

Architecture Our CNN architecture was found after 100 runs of the random search. It includes 4 convolutional blocks and 2 fully connected-layers. One convolutional block consists in one convolutional layer with a kernel size of 3, stride of 1 and padding of 1, an instance normalization layer, a leaky ReLU activation and a max pooling layer of kernel and stride of 2. A leaky ReLU activation was also inserted between the two fully-connected layers.

Training The network was trained to differentiate AD from CN participants by optimizing the cross-entropy loss criterion. The weights of the network were optimized using Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, no weight decay) with a learning rate of $3.31e-3$, during 100 epochs. A batch size of 8 is used. At the end of each epoch the loss of the network is computed on the validation set. The best network corresponds to the one which obtained the lowest validation loss value during the training process.

Computational setup We used a computing cluster of 20 nodes. Each node is composed of 4 Nvidia Tesla V100 32G GPUs and 24 cores. Each run required 10G and 1GPU during approximately 5h30 (this is an average time over all the runs in the random search, larger architectures requiring more time than smaller ones). Our final architecture required precisely 5h21 to be fully trained. The explainability method required 2h54 of running time on 1 GPU with 10G of memory footprint.

3.3 Evaluation strategy

Validation procedure We split the data set in three sets: training (75%), validation (10%) and test (15%). We ensured that the sex, age and label distribution was the same in all sets. This split was performed before the random search, and the best hyperparameter setup was chosen based on the result on the validation set only. The test set was used to estimate the performance of the chosen network and to generate the attribution maps.

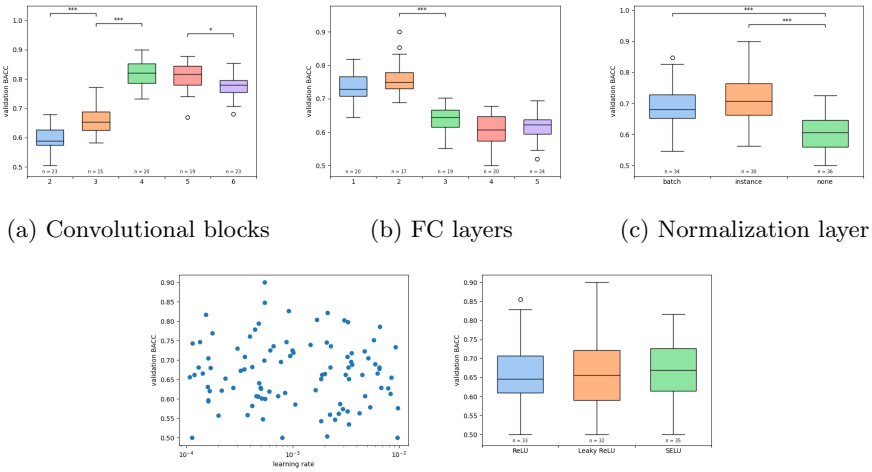
Metrics We evaluated the performance of the network by computing the sensitivity, specificity and balanced accuracy of the network.

Explainability The best network was explained by generating attribution maps with an occlusion method. This method consists in replacing each location of the input image and the its neighbouring voxels in a 5x5 cube by grey values (intensity=0.5). For each different location the loss of the network is assessed, then the value associated with this voxel corresponds to the absolute difference between this loss and the original loss (the one obtained without perturbing the image). This method allows to compute one attribution map per participant. We then computed the mean attribution map over the AD patients to get one attribution map representative of the disease at the data set level.

4 Results

The best network achieved a balanced accuracy of 92.3%, sensitivity of 100% and a specificity of 84.6% on the test set. The network failed at recognizing the oldest CN participants, the mean age of the badly classified CN participants being 78.6 years, whereas the correctly classified participants had a mean age of 73.5 years. No differences between these two groups in terms of sex, education and clinical scores were detected.

Hyperparameter search We studied how sensitive our network was to the different hyperparameter sets during the random search (see Figure 1). We observed that below 4 convolutional blocks the performance is significantly worse than with 4 or more. The performance also largely depends on the number of FC layers, and is degraded by a too large number of layers. The two normalisation layers give similar results, but significantly different from not using any normalisation process. The activation layers and different learning rate values gave similar results.



(a) Convolutional blocks (b) FC layers (c) Normalization layer

(d) Learning rate (e) Activation layer

Fig. 1: Balanced accuracy obtained on the validation set by the 100 runs of the random search depending on the different values of the hyperparameters. The significance of the difference between groups is evaluated with the Mann-Whitney U-test: * correspond to a p-value < 0.05, *** corresponds to a p-value < 0.001

Explainability The attribution maps of our network are displayed in Figure 2. Based on the AAL2 neuroanatomical parcellation, we quantified the mean value of the intensities in each region, normalised by its volume. The results for the top 5 regions can be found in Table 2.

5 Conclusion

In this article we showed that we could accurately differentiate demented patients from cognitively normal participants from their brain T1w-MRI. We found that our result depends on the chosen architecture, and especially the number of

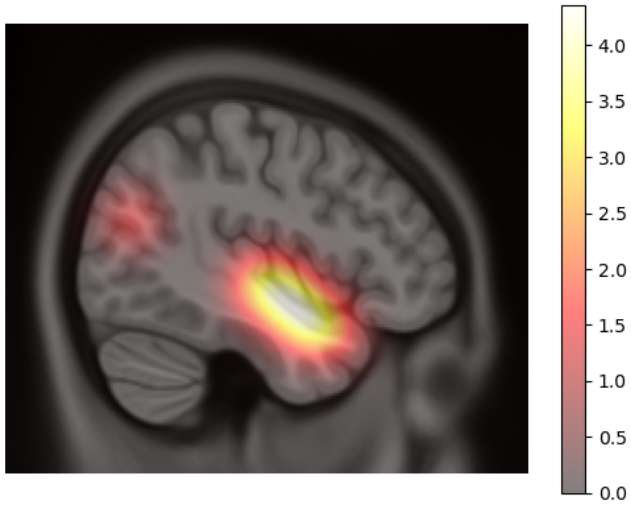


Fig. 2: Attribution map of the AD group superimposed to the template used for MRI preprocessing.

Table 2: Top 5 regions in which the mean normalised intensity was the highest in the attribution map.

Region	Side	Mean intensity	Max intensity
Hippocampus	left	0.93	2.01
Parahippocampal	left	0.92	4.56
Hippocampus	right	0.90	1.87
Amygdala	right	0.84	1.38
Temporal Superior pole	left	0.76	1.32

layers, but that the tendency was different for convolutional layers (a minimum of 4 were required) and the FC layers (the performance was hurt by adding more layers). The regions found in the attribution map corresponds to the ones clinically known to be affected in Alzheimer's disease, thus indicating that this network has the potential to be used in a clinical routine as a support tool for radiologists. Future work will investigate the robustness of the attribution maps towards the network hyperparameters.

References

1. Ewers, M., Sperling, R.A., Klunk, W.E., Weiner, M.W., Hampel, H.: Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. Trends in Neurosciences **34**(8) (2011) 430–442
2. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O.: Convolutional neural

networks for classification of Alzheimer's disease: Overview and reproducible evaluation. Medical Image Analysis **63** (2020) 101694
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet Classification with Deep Convolutional Neural Networks. In: Advances in neural information processing systems. Volume 25. (2012) 1097–1105
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015) 436–444
5. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6) (2010) 1310–1320
6. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis **12**(1) (2008) 26–41
7. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Brain Development Cooperative Group: Unbiased average age-appropriate atlases for pediatric studies. Neuroimage **54**(1) (2011) 313–327
8. Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L.: Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage **Supplement 1**(47) (2009) S102
9. Bergstra, J., Bengio, Y.: Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research **13**(Feb) (2012) 281–305