Generalization study of a deep learning classifier for the diagnosis of Alzheimer's disease

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. The ADNI data set has been widely used to train and evaluate deep learning networks detecting Alzheimer's disease from T1-weighted MRI. However, most studies do not guarantee a generalization of these networks to other data sets, which is crucial before considering a clinical application. In this study, we applied such CNN trained in a previous study to a subset of OASIS-1, and looked for the possible correlation between the network output and clinical scores, age and sex. We observed that the age leads to the strongest correlation, and that the sex is not correlated to the result. This questions the validity of this network.

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].

Numerous methods have been proposed but a majority has been developed and applied using images from the same dataset, often the Alzheimer's Disease Neuroimaging Initiative (ADNI), and does not assess the generalisability to other

Table 1: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores. Clinical scores were not provided for younger participants. Values are presented as mean (standard deviation) [range]. M: male, F: female

	Subjects	Age	Gender	MMSE	CDR
CN (old)	10	78.8 (9.4) [62, 90]	5F / 5M	28.8 (1.5) [25, 30]	0: 10
CN (young)	10	28.8 (12.9) [18, 55]	5F / 5M	–	–
AD	10	79.0 (6.3) [69, 88]	5F / 5M	26.5 (1.9) [23, 29]	0.5: 7; 1: 3

datasets. In this paper, we propose to perform such generalisability study by applying a model obtained by training on images from ADNI to images of the Open Access Series of Imaging Studies (OASIS)-1 dataset. We investigated if there was a correlation between the output of the network and demographic factors (age and sex) and clinical scores.

2 Data set

Data used in this work were obtained from the OASIS-1 data set[3], which consists of a cross-sectional collection of 416 subjects aged 18 to 96 comprising participants both cognitively normal (CN) and clinically diagnosed with very mild to moderate Alzheimer's disease (AD) [5]. The CN group was split into two groups depending on age: in the *old* group participants have a minimum age of 62 years, which corresponds to the minimal age of AD participants in this data set; the *young* group contains all participants strictly younger than 62. For each subject, among the multiple T1-weighted MR images available, we selected the average of the motion-corrected co-registered individual images resampled to 1 mm isotropic voxels, located in the `PROCESSED/MPRAGE/ SUBJ_111` subfolder. After the preprocessing pipeline, we randomly selected 10 participants in each group whose image passed the quality check procedure (see section 3.1). The cohort is further described in Table 1.

3 Methods

3.1 Preprocessing of T1-weighted MRI

The OASIS data have been curated and converted to the Brain Imaging Data Structure (BIDS) format [6] using Clinica [7,8] (v0.7.5). The T1-weighted MR images were pre-processed using the `t1-linear` pipeline of Clinica [7], which is a wrapper of the ANTs software [9]. Bias field correction was applied using the N4ITK method [10]. An affine registration to MNI space was performed using ANTs [11]. The registered images were further rescaled based on the min

[3] OASIS: `https://oasis-brains.org`

and max intensity values. Images were then cropped to remove the background resulting in images of size 169×208×179, with 1 mm isotropic voxels. To ensure the reproducibility of the result, the random seed of ANTs was set to 42.

We performed quality control on the outputs of the preprocessing procedure using the DL-based framework proposed by Fonov et al. [12] and implemented in ClinicaDL [13] (v1.4.0). This software outputs a probability indicating how accurate the registration is. We excluded the scans with a probability lower than 0.5 and visually checked the remaining scans whose probabilities were lower than 0.70. As a result, 39 scans were excluded.

3.2 Deep learning network

Hyperparameter search & values We reused one of the networks trained on the ADNI dataset in a previous study [2]. More precisely, we used experiment 3 as listed in the supplementary Table 4 of the original paper. In this experiment, the network was trained on the full images preprocessed as described above. First an auto-encoder was trained to reproduce all available baseline images (AD, CN and MCI). Then a CNN was built by reusing the encoder part and adding the fully-connected layers. Finally, this CNN learned to differentiate AD patients from CN participants from baseline sessions only. The architecture of the network consists of five convolutional and three fully connected layers and is displayed in Figure 1. Other architecture and hyperparameter exploration details can be found in the open-source original paper [2].

Computational setup The application of the CNN does not require extensive computational resources and could run without a GPU in 5 minutes on a Apple M1 Pro. The maximum memory use was 7G.

3.3 Evaluation strategy

Metrics The performance of the network was evaluated by computing the confusion matrix of the binary classification task. The strength of the correlation between the probability of the CN class computed by the network and other factors was computed as follows with the Spearman correlation coefficient and associated p-value.

Note that clinical scores were not provided for the young CN group, so we assumed that they obtained the best possible scores: a MMSE score of 30 and a CDR score of 0.

Interpretability Though many explainability methods exist, we restricted this study to attribution maps produced by gradient back-propagation [14], as it is widely used and conceptually simple. An individual attribution map corresponds to the gradients of an output node with respect to an image. In our case, the output node is the one corresponding to the CN group. The intensities of the attribution map of an image correspond to the changes needed to transform this image into a sample of the CN group. We computed the group attribution map

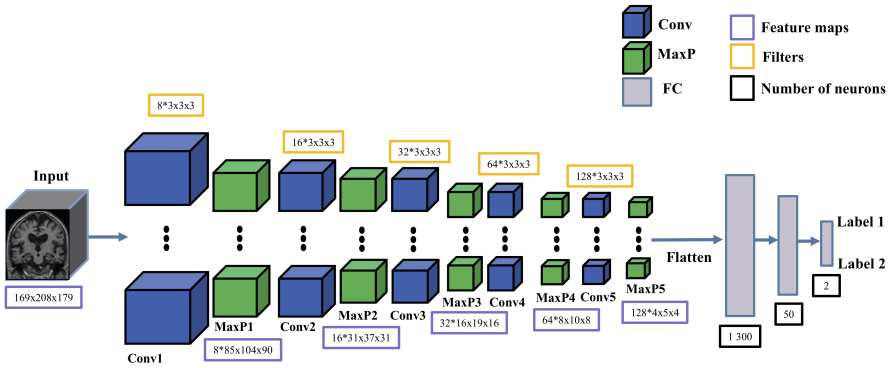

Fig. 1: Architecture of the 3D subject-level CNN. For each convolutional block, we only display the convolutional and max pooling layers. Filters for each convolutional layer represent the number of filters * filter size. Feature maps of each convolutional block represent the number of feature maps * size of each feature map. Conv: convolutional layer; MaxP: max pooling layer; FC: fully connected layer. Figure reproduced from [2].

of AD patients, corresponding to the mean value of the 10 attribution maps of the AD patients of our data set. This method computes very noisy outputs as it is voxel-based, hence to better visualize the regions, we applied a Gaussian filter of standard deviation $\sigma = 2$ to the group attribution map.

4 Results

The performance of the network is lower when considering only the old population compared to using the whole CN group. Indeed 8 old CN participants on 10 are classified as AD patients (see Table 2).

Table 2: Confusion matrix of the CNN.

	AD	CN
AD	9	1
CN (old)	8	2
CN (young)	0	10

Both clinical scores are correlated with the probability of the diagnosis. However the strongest correlation is with age (correlation coefficient = 0.87). The sex is not correlated with the diagnosis.

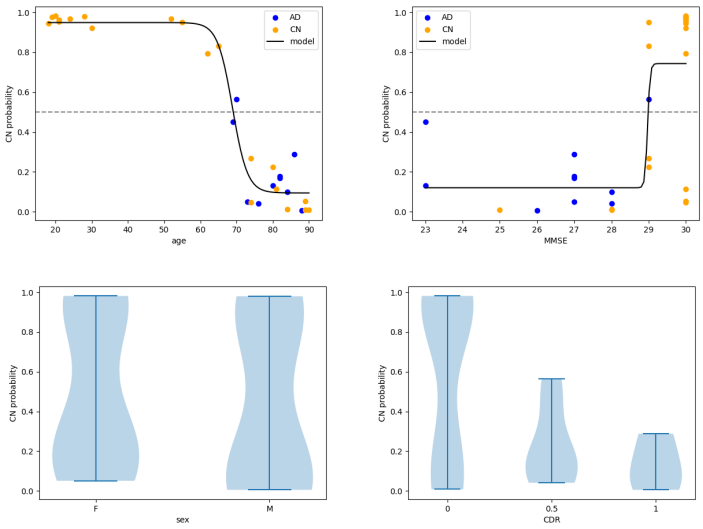

Fig. 2: Correlation with the probability of the CN class computed by the network. For continuous variables (a, b) a logistic model was fitted to the data. For categorical variables (c, d) the correlation is illustrated with violin plots. For each case the correlation coefficient (CC) and the p-value of the Spearman test are added in the caption between brackets.

On the attribution map (Figure 3) we see that the medial temporal lobe, which is known to be atrophied in Alzheimer's disease, is highlighted. However on the central slices (75 & 95) the network also focuses on regions outside the brain, and also next to the cerebellum.

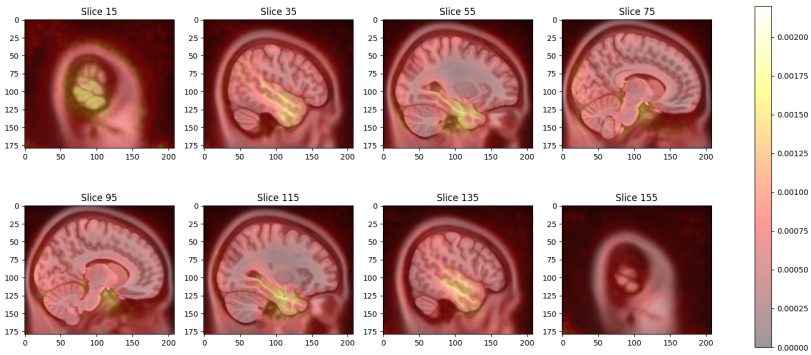

Fig. 3: AD group attribution map of the network trained on the first fold of the 5-fold cross-validation, superimposed on the cropped template used to preprocess the images (ICBM 2009c Nonlinear Symmetric)

5 Conclusion

In this paper we showed that the application of a deep learning network trained to detect AD on the ADNI data set cannot be directly applied to OASIS-1. Though the result of the network is correlated to the clinical scores used to diagnose dementia (MMSE and CDR), and the attribution map is highlighting regions that are known to be affected by the disease, we found that the strongest correlation is with the age, which is a healthy cause for brain atrophy. Future work will investigate why other regions than the medial temporal ones are included in the attribution map, and how they could be correlated with age detection.

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. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience **19**(9) (2007) 1498–1507
6. Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Flandin, G., Ghosh, S.S., Glatard, T., Halchenko, Y.O., Handwerker, D.A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B.N., Nichols, T.E., Pellman, J., Poline, J.B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J.A., Varoquaux, G., Poldrack, R.A.: The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data **3** (2016) 160044
7. Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.O., Durrleman, S., Colliot, O.: Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. Frontiers in Neuroinformatics **15** (2021) 39
8. Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.O., Durrleman, S., Evgeniou, T., Colliot, O.: Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. NeuroImage **183** (2018) 504–521

9. Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C.: The Insight ToolKit image registration framework. Frontiers in Neuroinformatics **8** (2014) 44
10. 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
11. 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
12. Fonov, V.S., Dadar, M., Adni, T.P.A.R.G., Collins, D.L.: Darq: Deep learning of quality control for stereotaxic registration of human brain mri to the t1w mni-icbm 152 template. NeuroImage **257** (2022) 119266
13. Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., Burgos, N.: ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing. Computer Methods and Programs in Biomedicine **220** (2022) 106818
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: In Workshop at International Conference on Learning Representations. (2014)