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 low level of accuracy in distinguishing AD patients from cognitively normal controls. On the test set, the model achieved a balanced accuracy of 71.3%. The attribution maps associated to the trained network showed different patterns after retraining the networks.
This reproducible study questions the potential of convolutional neural networks in effectively differentiating Alzheimer's disease patients from cognitively normal individuals based on brain MRI scans. 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

The data set used to train the network is OASIS-3[3][5]. This data set includes 755 cognitively normal adults and 622 individuals at various stages of cognitive decline ranging in age from 42 to 95 years, as described in Table 1.

For our study we defined two different labels, which mainly depend on the value of the Clinical Dementia Rating (CDR) score:

- **CN** includes all the images of the cognitively normal adults (CDR=0) with a valid T1-MRI at baseline.
- **AD** includes all the images of the individuals with mild dementia (CDR>=1) with a valid T1-MRI at baseline.

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

Participants	Images	Age	Gender	MMSE	CDR
CN 723	1245	73.4 (5.9) [45, 89]	354 M / 369 F	29.1 (1.1) [24, 30]	0: 723
AD 543	874	75.6 (6.9) [50, 92]	245 M / 298 F	23.2 (2.1) [18, 27]	1: 368; 2: 175

The images were considered as valid if they passed an automatic quality check [6]. This procedure relies on a pre-trained deep learning network[4] which 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 probability were lower than 0.70. As a result, 24 CN participants and 32 AD patients were removed.

[3] https://www.oasis-brains.org
[4] https://github.com/vfonov/deep-qc

3 Methods

3.1 Preprocessing of T1-weighted MRI

For anatomical T1w MRI, the preprocessing pipeline was based on SPM12. First, the Unified Segmentation procedure [7] is used to simultaneously perform tissue segmentation, bias correction and spatial normalization of the input image. Next, a group template is created using DARTEL, an algorithm for diffeomorphic image registration [8], from the subjects' tissue probability maps on the native space, usually GM, WM and CSF tissues, obtained at the previous step. Here, not only the group template is obtained, but also the deformation fields from each subject's native space into the DARTEL template space. Lastly, the DARTEL to MNI method [8] is applied, providing a registration of the native space images into the MNI space: for a given subject its flow field into the DARTEL template is combined with the transformation of the DARTEL template into MNI space, and the resulting transformation is applied to the subject's different tissue maps. As a result, all the images are in a common space, providing a voxel-wise correspondence across subjects. They all have the same size of 145x145x145.

3.2 Deep learning network

Hyperparameter search We looked for the best training hyperparameters, and then we ran a grid search including:

- the learning rate, which could take 5 different values (10e-1, 10e-2, 10e-3, 10e-4, 10e-5),
- the weight decay, which could take 2 different values (0, 0.01),
- the optimization algorithm (Adam, Adagrad or Adadelta).

We chose the set of hyperparameters leading to the best performance. During the grid search all networks were initialised with the same weights and the train/validation split was identical for all runs.

Architecture We used a common deep learning architecture: ResNet [9]. This architecture takes as input 2D images, thus for each volume we extracted the middle axial slice of the 3D volume (z=72) and gave it as input to the network.

Training All networks were trained during 100 epochs, to minimize the cross-entropy loss. The hyperparameter search led to the choice of the Adagrad optimizer, with a weight decay of 0.01 and a learning rate of 10e-3. The weights of the network were updated based on batches of 32 images. During training, we used the 5 neighbouring slices to the middle one (z=70-74) to perform data augmentation.

Computational setup All experiments ran on a computer with 24 cores and one NVIDIA GPU (GeForce GTX 1650). Each run of the training task required 4h20. The evaluation and explainability tasks only needed a few seconds.

3.3 Evaluation strategy

Validation procedure For each label, 100 volumes were randomly chosen for evaluation and explainability purposes. The rest of the images were split between training (80%) and validation (20%) sets. For some experiments this split was kept identical by setting a value for the random seed used.

Explainability We explained our networks by generating attribution maps with the gradient back-propagation method [10]. Once the network is trained, each image of the CN group was propagated through the network, then the gradients corresponding to an increase of the value of the node associated to the AD class were back-propagated to the level of the image. These gradients represent how these CN images should change so the network classifies them in the AD class. We assumed that the value of the absolute value of the gradient was correlated with the importance of each pixel, and that the sign didn't carry any information on the importance of the pixel, then we used the absolute value of the mean map obtained over all CN images in the test set. AD images were not used for explainability purposes.

Metrics For each experiment 30 networks were trained. We compared the variability of the balanced accuracy (BACC) of these networks for each experiment. To evaluate the robustness of the explainability method we used the mean squared error (MSE) between each map obtained with a network of the experiment and the average map obtained with all the networks of this experiment.

4 Results

We evaluated the robustness of our setup by evaluating the variability of the result if (1) the initial weights are identical, (2) the train/validation splits are identical, (3) both of the 2 previous conditions are fulfilled, (4) none of them are fulfilled.

Training robustness All the different setups gave similar mean performance. We observe in Figure 1 a significance in the variability obtained in the setup (3), with both the initial weights and the train/validation splits being fixed, and all the other setups. The other setups, in which only one condition or none was fixed showed non-significantly different variabilities in the network's success.

Explainability robustness For all the setups the variability around the mean CN attribution map over the 30 runs was similar (Figure 2). Fixing the initial weights and/or the train/validation splits didn't seem to lower the variability of the attribution maps obtained. The mean CN attribution maps obtained for the first run of each setup are shown in Figure 3.

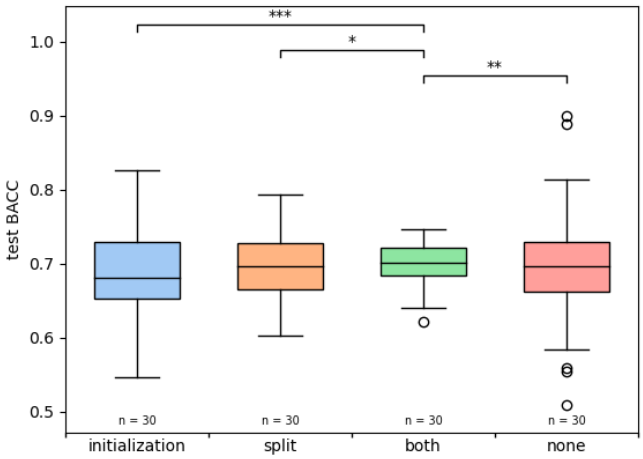


Fig. 1: Variability of the balanced accuracy on the test set obtained for 30 runs of each setup. (1) identical intialisation, (2) identical train/validation splits, (3) identical initialisation and train/validation splits, (4) no fixed initialisation or splits. The Levene test was used to compute the significance of the difference between the variability of the performance. : * correspond to a p-value < 0.05, ** corresponds to a p-value < 0.01, *** corresponds to a p-value < 0.001

5 Conclusion

Our network performed poorly with a balanced accuracy of 71.3%. We showed that the training process is not equally stable and that its variability can be lowered by fixing the initial weights of the network and the images used for training and validation. However, this is still not enough to guarantee a deterministic result as the learnt weights also depend on the order in which the batches of images are fed in the network during training. Another important conclusion is that this wasn't insufficient to guarantee the stability of the attribution maps, which remain equally different from run to run even when the initialisation and split are fixed. This variability questions the ability of the network to robustly identify patterns to predict Alzheimer's disease from T1-MRI.

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

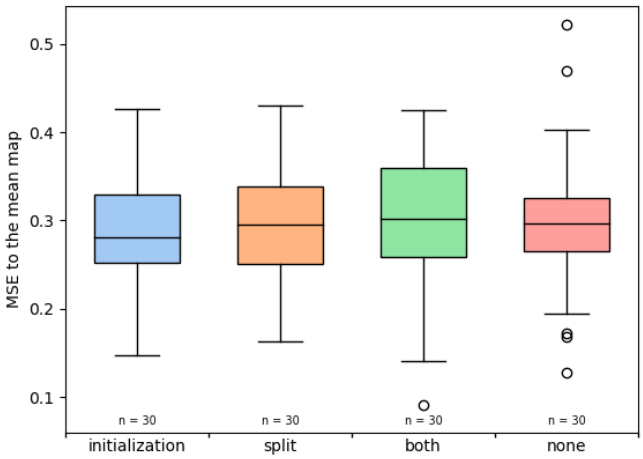
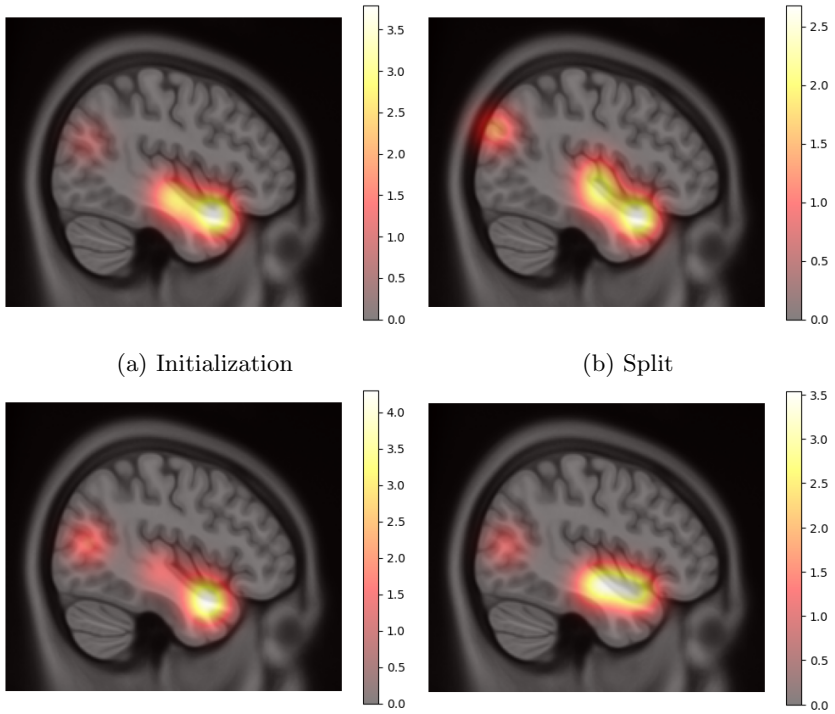


Fig. 2: Variability of the balanced accuracy on the test set obtained for 30 runs of each setup. (1) identical intialisation, (2) identical train/validation splits, (3) identical initialisation and train/validation splits, (4) no fixed initialisation or splits. The Levene test was used to compute the significance of the difference between the variability of the performance. : * correspond to a p-value < 0.05, ** corresponds to a p-value < 0.01, *** corresponds to a p-value < 0.001

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. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D.: OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease (2019)
6. Fonov, V.S., Dadar, M., Group, T.P.A.R., Collins, D.L.: Deep learning of quality control for stereotaxic registration of human brain MRI. bioRxiv (2018) 303487
7. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage **26**(3) (2005) 839–851
8. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1) (2007) 95–113
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition, IEEE (2016) 770–778 Comment: Tech report.
10. 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)



(a) Initialization (b) Split

(c) Both (d) None

Fig. 3: Attribution maps for the CN group obtained on the first run of each setup.