A Multi-modal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease

Conference Paper *in* Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference · July 2018

DOI: 10.1109/EMBC.2018.8512468

5 authors, including:

Simeon Spasov
Amazon Alexa
25 PUBLICATIONS **512** CITATIONS
SEE PROFILE

Luca Passamonti
Biogen
288 PUBLICATIONS **8,039** CITATIONS
SEE PROFILE

Pietro Lio
University of Cambridge
915 PUBLICATIONS **26,565** CITATIONS
SEE PROFILE

Nicola Toschi
University of Rome Tor Vergata
301 PUBLICATIONS **8,951** CITATIONS
SEE PROFILE

A Multi-modal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease

Simeon E. Spasov, Luca Passamonti, Andrea Duggento, Pietro Liò, and Nicola Toschi, Senior Member IEEE

Abstract— **This paper presents a multi-modal Alzheimer's disease (AD) classification framework based on a convolutional neural network (CNN) architecture. The devised model takes structural MRI, and clinical assessment and genetic (APOe4) measures as inputs. Our CNN structure is designed to be efficient in its use of parameters which reduces overfitting, computational complexity, memory requirements and speed of prototyping. This is achieved by factorising the convolutional layers in parallel streams which also enables the simultaneous extraction of high and low level feature representations. Our method consistently achieves high classification results in discriminating between AD and control subjects with an average of 99% accuracy, 98% sensitivity, 100% specificity and an AUC of 1 across all test folds. Our study confirms that careful tuning of CNN characteristics can result in a framework which delivers extremely accurate predictions in a clinical problem despite data paucity, opening new avenues for application to prediction tasks which regard patient stratification, prediction of clinical evolution and eventually personalised medicine applications.**

I. Introduction

A growing number of experimental studies using machine learning techniques have employed different approaches to categorise and classify Alzheimer's disease (AD), the most common form of dementia world-wide. The proliferation of these methods has been motivated by the fact that computer-aided classification of AD might offer a sophisticated, automated, and potentially objective mean to systematically discriminate amongst patients with dementia disorders, which ultimately will help diagnosis and prognostication, especially in clinically challenging contexts. A related promise of this conceptual framework regards the possibility of developing reliable and stable markers of brain pathology with high sensitivity and specificity in detecting single cases. This is a fundamental requirement in the arena of 'personalised medicine', a new field of research which aims at creating individualised patient-specific profiles for a precise assessment of the risk of developing neurodegenerative and other neurological disorders. These individualised profile can be based on several and usually multi-modal indices including genetic risk factors, life-style measures, demographic characteristics, and cognitive indices. The possibility of summarising these complex and multi-dimensional data into a single, meaningful score will therefore greatly facilitate clinical decision-making and will offer increased objectivity in supporting these decisions.

Many methods for computer-aided Alzheimer's disease diagnosis which use functional and structural image modalities, such as MRI, fMRI, DTI, PET, have been developed. A lot of earlier work focused on volumetric-based approaches applied on different scales of the image. Examples include morphometric methods, such as voxel-based morphometry (VBM) [1], which investigates voxel-wise differences in grey and white matter concentration between groups of subjects, tensor-based morphometry (TBM) [2] which identifies regional differences in the jacobian determinant of high-dimensional nonlinear transformations, and deformation-based morphometry (DBM) [3] which focuses on macroscopic discrepancies between populations by taking advantage of the full deformation field computed when normalising all brains into a common space. Inter-subject variability, however, prompted the development of feature-engineering approaches designed to extract descriptive factors from images and classify them using machine learning [4], [5], [6]. The advent of convolutional neural networks (CNNs) meant that the feature engineering aspect could be optimised by integrating it in the training procedure. One application is in AD vs healthy discrimination which produces state-of-the-art results [7].

This study describes a multi-modal CNN which uses a layer factorisation approach that reduces the number of network parameters while retaining network depth. The proposed model is applied on the AD/healthy classification problem. The data we utilise is structural MRI, and clinical assessment and genetic (APOe4, apolipoprotein E expression level) measures.

II. Materials and Methods

A. ADNI Data

The data we used in this work is from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI is a multicenter longitudinal study to identify imaging, clinical, genetic and biochemical biomarkers for the early detection and tracking of Alzheimer's disease. The initiative was funded by a $67 million partnership by the public and private sector. The initial phase known as ADNI1 included 200 normal controls (NC), 400 MCI and 200 AD subjects between 55-90 years of age from various sites from the US

*Research supported by the Engineering and Physical Sciences Research Council [EP/L015889/1].

S. Spasov is with the University of Cambridge, Cambridge, Department of Computer Science and Technology, UK (email: ses88@cam.ac.uk).
L. Passamonti is a Senior Clinical Research Fellow at the Department of Clinical Neurosciences , University of Cambridge, Cambridge (e-mail: lp337@medschl.cam.ac.uk).
A. Duggento is a Research Fellow at the University of Rome "Tor Vergata", Department of Biomedicine and Prevention, Italy (email: duggento@med.uniroma2.it)
P. Liò is an Associate Professor at the University of Cambridge, Cambridge, Department of Computer Science and Technology, UK (email: pl219@cam.ac.uk).
N. Toschi is an Associate Professor in Medical Physics at the University of Rome "Tor Vergata", Department of Biomedicine and Prevention, and a Visiting Professor and assistant in Neurology at the Athinoula A. Martinos Center for Biomedical Imaging at Harvard Medical School (e-mail: toschi@med.uniroma2.it).

978-1-5386-3646-6/18/$31.00 ©2018 IEEE

TABLE I

CLINICAL MEASURES AND DEMOGRAPHIC CHARACTERISTICS OF HEALTHY SUBJECTS AND AD PATIENTS

	No. of subjects	Age (years)	Male/ Female	years in education	APOe4 expression level			CDRSB	ADAS11	ADAS13	RAVLT		
					0	1	2				immediate	learning	forgetting
AD	**192**	**75.6±7.5**	**103/81**	**15±2.9**	**57**	**86**	**41**	**4.4±1.6**	**18.8±6**	**29±7.3**	**23±7**	**1.7±1.8**	**4.4±1.9**
NC	**184**	**74.6±6**	**92/100**	**16.3±2.7**	**144**	**43**	**5**	**0.2±0.9**	**6±3.8**	**9.3±5.7**	**44±10.5**	**6±2.4**	**3.7±2.7**

and Canada. It is extended with ADNI GO and ADNI2 which add new participants and funding to the study. The database is made available to researchers around the world. The principal investigator of ADNI, who oversees all aspects, is Dr. Michael Weiner, MD, VA Medical Center and University of California - San Francisco. For up-to-date information, see www.adni-info.org.

Data used in this study was obtained from ADNI1 and comprises MRI (T1- weighted MP-RAGE images at 1.5T with 256 x 256 x 256 voxels and a target voxel size of 1 mm^3), and clinical features (demographic data: age, gender, ethnic and racial categories, years in education; biofluids: APOe4 genotyping; behavioural assessment: clinical dementia rating - CDRSB; Alzheimer's disease assessment scale - ADAS11, ADAS13; and the episodic memory evaluations in the Rey Auditory Verbal Learning Test (RAVLT_immediate, RAVLT_learning, RAVLT_forgetting, RAVLT_perc_forgetting). The subjects were grouped as (i) AD if they had been diagnosed as having Alzheimer's at baseline (n = 192) ; (ii) NC (normal controls) if diagnosis was normal at baseline and subjects did not convert to AD within the next three years of periodic clinical and imaging followup (n = 184).

B. *Data Pre-processing*

All T1-weighted scans were normalised to the Montreal Neurological T1 Template (MNI152) using the Advanced Normalization Tool, ANTs. [8] The high-dimensional non-linear transformation (symmetric diffeomorphic normalisation transformation) model was initialised through a generic linear transformation which consisted of center of mass alignment, rigid, similarity and fully affine transformations followed by (metric: neighbourhood cross correlation, sampling: regular, gradient step size: 0.12, four multi-resolution levels, smoothing sigmas: 3, 2, 1, 0 voxels in the reference image space, shrink factors: 6, 4, 2, 1 voxels. We also used histogram matching of images before registration and data winsorisation with quantiles: 0.001, 0.999. The convergence criterion was set to be as follows: slope of the normalised energy profile over the last 10 iterations < 10-8). Coregistration of all scans required approximately 19200 hour of CPU time on a high-performance parallel computing cluster. All images in MNI space were therefore comprised of a total of 182 x 218 x 182 = 7.22 million voxels per image. After normalisation, all images were masked to include only brain tissue.

Numerical normalisation for the pre-processed MRI images is performed per sample - each 3D volume is standardised to 0 mean and unit standard deviation. The reasoning is that brain atrophy can be recognised as an in-sample shift in intensity for a certain area compared to other regions. Clinical feature normalisation also follows the feature scaling procedure, where the values of each separate clinical factor are normalised between [0, 1] .

C. *Neural Network Architecture*

The major advantage of deep learning in computer vision is the ability to automatically extract features without the need for laborious manual feature engineering. The cost, however, is the high number of parameters required by all the network layers which is known to increase the tendency of overfitting to unseen samples. One way to remedy this issue is by introducing more training examples, which is impractical when dealing with medical datasets owing to the cost of performing physical examinations and clinical tests on patients. Thus, it is left at the discretion of the deep learning practitioner to develop an architecture which uses parameters very efficiently. The architecture of the network we devised is summarised in fig. 2. The topology extracts lower dimensional feature embeddings from two separate input sources - MRI images and clinical assessment measures,

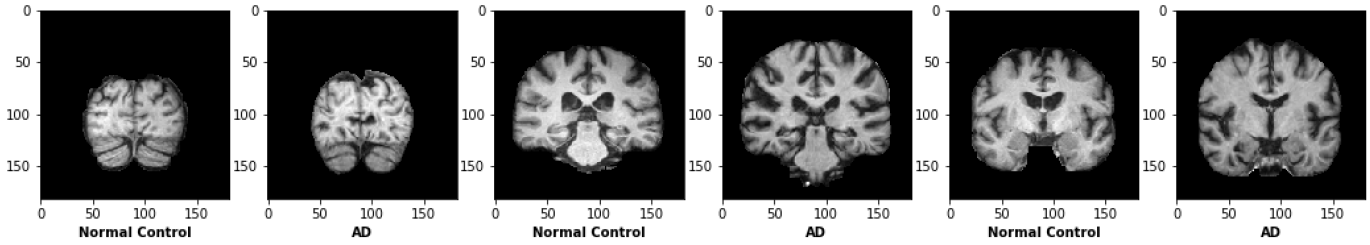

Figure 1. Examples of three pairs of slices from the normalised MRI volumes we use for Alzheimer's patients and normal controls.

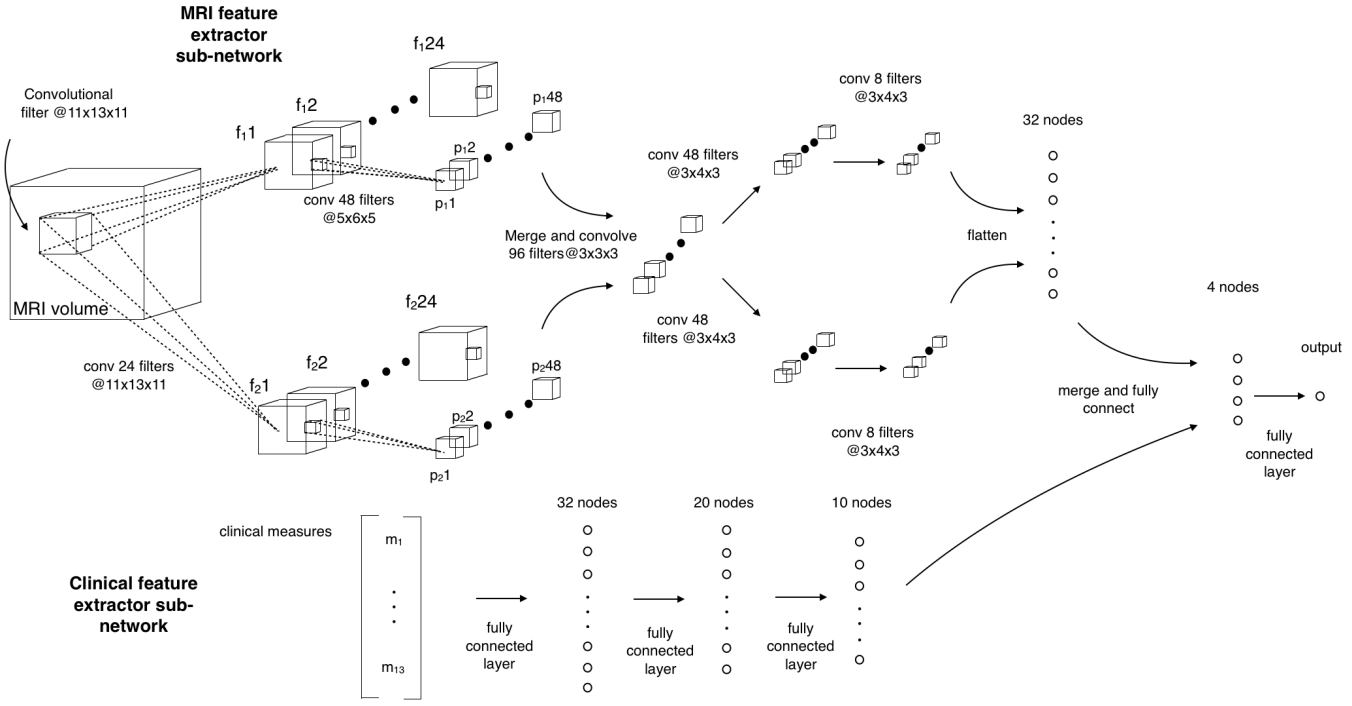

Figure 2. An illustration of the architecture of our neural network. It receives two types of inputs: a 13-dimensional vector (**m**) of clinical measures and a 182x218x182 MRI volume. Descriptive features are extracted from these inputs by two *sub-networks*. The MRI feature extractor is a 3D CNN - the dashed lines depict how a single feature at a given level of the network is produced via a transformation of its receptive field in the previous layer. Sliding this filter with certain strides in all dimensions produces the feature maps for a single channel in the next hierarchical level. The solid arrows represent the operations we apply to each feature map. The feature maps in the first two layers are denoted by **f** and **p**, where the subscript denotes the convolutional stream. The first layer has 24 channels in each stream, whereas the second layer - 48 channels.

before merging and passing them through a final fully connected layer with a single output between 0 and 1.

The MRI feature extractor sub-network consists of 5 sequential convolutional layers followed by a fully connected layer. Some of the convolutional layers are split in two before convolving them at the next stage so as to reduce the number of parameters in the latter layer. As means of visualisation, one could imagine we had adhered to the standard approach of having two consecutive convolutional layers with F1 and F2 channels respectively. Consequently, we would have an overall of $(k_1k_2k_3 + 1)F_1F_2$ weights in the latter layer, where $k_{1\ldots3}$ are the dimensions of its kernel. On the other hand, if the second layer accepts only half of the channels from the previous stage, the overall number of parameters is halved. The network loses some expressive power since the latter layer would only accept part of the feature maps from the previous layer, nevertheless, this comes with the flexibility of having different kernel sizes and strides in the two streams. Hence, each layer can extract fine-grained details *and* more abstract, "big picture" representations simultaneously. This idea is explored in greater detail in GoogleNet [14].

In theory, this factorisation procedure can be repeated until we are left with a single channel. However, the depth of the network would also increase which could introduce training difficulties. Hence, we opted for a simpler dual-stream architecture. The streams are merged and convolved to achieve information exchange before diverging again. The dual-stream architecture achieves an overall parameter count of ~820,000 versus ~1.2 million for an equivalent sequential network with the same number of filters. The model was implemented in Keras with a TensorFlow backend and trained on a single Nvidia Pascal TITAN X GPU. The first two convolutional layers of the MRI feature extractor employ 24 and 48 filters respectively in each stream with kernel sizes of 11x13x11 and 5x6x5 and strides of 4 and 1 voxels in all dimensions. After merging these representations we extract 48 channels with kernels of size 3x4x3. The output channels are then factorised in two to form the second dual stream of 3x4x3 convolutions. Finally, we concatenate and flatten the feature maps to pass them through a fully connected layer of 32 hidden. The stride length of all convolutional layers but the first is set at 1x1x1 voxels. The spatial size of the feature maps is reduced via max pooling which is applied after every layer with strides of 2 voxels and a receptive field of 3 voxels. All of the layers have exponential linear unit (ELU) activations.

The clinical feature extractor is a series of three fully connected layers with 40, 20 and 10 hidden units respectively with ELU activations.

Standard deep learning techniques to reduce overfitting were also utilised. We used batch normalisation [10] - a technique which normalises the outputs of each layer, thus accelerating the rate of training and acting as a regulariser; dropout [11] which works by randomly dropping units and their connections during training; and l2 regularisation which penalises weights of high absolute value, hence directly limiting the variety of functions our model can represent.

During training we minimised the binary cross entropy between the model's predictions and the true labels for Alzheimer's and healthy subjects. We used the Adam optimiser [12] with a linearly decaying learning rate with an initial value of 5e-4 which reaches 0 at the end of training; the β1, β2 and ε were kept at their default values provided in the original Adam paper. Dropout was set at 0.02, the l2 regularisation hyper parameter was 10^{-5}.

D. Training and Evaluation Procedure

For the evaluation of the classifier, we repeated the same sampling strategy to divide the samples in training, validation and test set splits with 308/36/32 samples respectively. Both the validation and test sets were balanced between the two classes and were formed by randomly picking an equal number of healthy and AD subjects. Finally, the remaining samples comprised the training set. The model was trained for 25 epochs and then evaluated on the test set. The validation set was used to ensure the model was not overfitting. This procedure was then repeated 10 times with different sampling seeds so as to minimise the effect of random variation. The evaluation metrics used are accuracy (ACC: the percentage of correctly labelled samples), SEN (sensitivity: the number of correctly classified AD patients divided by total number of AD subjects), SPE (specificity: the number of correctly classified heathy patients divided by total number of healthy subjects), PPV (positive predictive value: the number of AD subjects divided by the number of AD predictions), NPV (negative predictive value: the number of healthy subjects divided by the number of NC predictions).

III. Results

Our model achieved an AUC of 1 with an optimal operating point accuracy of 99% on average. The majority of the times (6 out of 10) the network would predict with no misclassification. The specificity of the method is 100% across all testing folds. The non-perfect scores are achieved due to a misclassification of a single Alzheimer's subject, hence decreasing average sensitivity to 98%. We have found 8 outliers (out of 376) in the normal controls who exhibit feature characteristics more typical of the AD group. Hence, this would explain the inability of the approach to consistently predict with 100% accuracy. The model's relatively low parameter count compared to other CNNs, such as AlexNet [9], VGGNet [13], makes it more suitable for data-scarce clinical studies. It is less prone to overfitting and quick to fine-tune and prototype as it takes ~20-30 sec./ epoch on an Nvidia PASCAL TITAN X GPU.

V. Discussion and Conclusions

This study shows that our framework was able to successfully discriminate between patients with Alzheimer's disease and healthy controls with almost perfect accuracy.

Table II

Model performance in discriminating AD patients from healthy controls on the test set

	AUC	ACC	SEN	SPE	PPV	NPV
@ Optimal Operating Point	**1**	**99%**	**98%**	**100%**	**100%**	**98.2%**

Overall, these findings represent an important 'sanity check' of the ability of the algorithm to deal with multi-modal data including brain scanning as well as clinical continuous and categorical variables. Additional work is now in progress to address diagnostic challenges of highly related clinical conditions; for example, the discrimination between people with forms of mild cognitive impairment (MCI) who go on and develop AD relative to individuals with benign types of MCI who do not develop AD. Further applications of the algorithm regard the potential to predict clinically useful outcome and prognostic measures including for instance the time to conversion from MCI to AD as well as the implementation of dimensionality reduction strategies in order to extract the true degrees of freedom which underlie the still not well understood AD spectrum as well as its representation in multimodal clinical/imaging features. Additional implementations to include extra layers of information (for example genetic data, blood tests, other imaging modalities) are also warranted. Together, these refinements of the algorithm will help create individualised profile risks to successfully predict the development of AD in single cases, which will have a major impact in personalised therapeutic scenarios.

References

1. J. Ashburner and K. J. Friston, "Voxel-Based Morphometry—The Methods," NeuroImage, vol. 11, no. 6, pp. 805–821, Jun. 2000.
2. Colin Studholme, Corina Drapaca, Bistra Iordanova, and Valerie Cardenas, "Deformation-based mapping of volume change from serial brain mri in the presence of local tissue contrast change," IEEE transactions on Medical Imaging, 25(5):626–639, 2006.
3. John Ashburner,* Chloe Hutton, Richard Frackowiak, Ingrid Johnsrude, Cathy Price, and Karl Friston, "Identifying Global Anatomical Differences: Deformation-Based Morphometry," Human Brain Mapping 6:348–357(1998)
4. X. Long, L. Chen, C. Jiang, and L. Zhang, "Prediction and classification of Alzheimer disease based on quantification of MRI deformation," PLOS ONE, vol. 12, no. 3, p. e0173372, Mar. 2017.
5. X. Liu, D. Tosun, M. W. Weiner, and N. Schuff, "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification," NeuroImage, vol. 83, pp. 148–157, Dec. 2013.
6. A. Chincarini et al., "Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease," NeuroImage, vol. 58, no. 2, pp. 469–480, Sep. 2011.
7. S. Sarraf and G. Tofighi, "Deep learning-based pipeline to recognize Alzheimer's disease using fMRI data," in 2016 Future Technologies Conference (FTC), 2016.
8. B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," NeuroImage, vol. 54, no. 3, pp. 2033–2044, Feb. 2011.
9. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks.," in NIPS, 2012, pp. 1106–1114.
10. S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.," in ICML, 2015, vol. 37, pp. 448–456.
11. N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
12. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization.," CoRR, vol. abs/1412.6980, 2014.
13. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition.," CoRR, vol. abs/1409.1556, 2014.
14. C. Szegedy et al., "Going deeper with convolutions.," in CVPR, 2015, pp. 1–9.