NeuroImage 189 (2019) 276–287

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease

Simeon Spasov [a,*], Luca Passamonti [b], Andrea Duggento [c], Pietro Liò [a,1], Nicola Toschi [c,d,1], for the Alzheimer's Disease Neuroimaging Initiative[2]

[a] University of Cambridge, Cambridge, Department of Computer Science and Technology, William Gates Building, 15 J J Thomson Ave, Cambridge, CB3 0FD, UK
[b] Department of Clinical Neurosciences, University of Cambridge, Herchel Smith Building, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SZ, Cambridge, UK
[c] Department of Biomedicine and Prevention, University of Rome "Tor Vergata", Via Cracovia, 00133, Roma, RM, Italy
[d] A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, USA

ARTICLE INFO

Keywords:
Deep learning
Neural networks
Classification
Mild cognitive impairment
Alzheimer's disease
Magnetic resonance imaging
ADNI
Early diagnosis

ABSTRACT

Some forms of mild cognitive impairment (MCI) are the clinical precursors of Alzheimer's disease (AD), while other MCI types tend to remain stable over-time and do not progress to AD. To identify and choose effective and personalized strategies to prevent or slow the progression of AD, we need to develop objective measures that are able to discriminate the MCI patients who are at risk of AD from those MCI patients who have less risk to develop AD. Here, we present a novel deep learning architecture, based on dual learning and an *ad hoc* layer for 3D separable convolutions, which aims at identifying MCI patients who have a high likelihood of developing AD within 3 years.

Our deep learning procedures combine structural magnetic resonance imaging (MRI), demographic, neuropsychological, and APOe4 genetic data as input measures. The most novel characteristics of our machine learning model compared to previous ones are the following: 1) our deep learning model is multi-tasking, in the sense that it jointly learns to simultaneously predict both MCI to AD conversion as well as AD vs. healthy controls classification, which facilitates relevant feature extraction for AD prognostication; 2) the neural network classifier employs fewer parameters than other deep learning architectures which significantly limits data-overfitting (we use ~550,000 network parameters, which is orders of magnitude lower than other network designs); 3) both structural MRI images and their warp field characteristics, which quantify local volumetric changes in relation to the MRI template, were used as separate input streams to extract as much information as possible from the MRI data. All analyses were performed on a subset of the database made publicly available via the Alzheimer's Disease Neuroimaging Initiative (ADNI), (n = 785 participants, n = 192 AD patients, n = 409 MCI patients (including both MCI patients who convert to AD and MCI patients who do not covert to AD), and n = 184 healthy controls).

The most predictive combination of inputs were the structural MRI images and the demographic, neuropsychological, and APOe4 data. In contrast, the warp field metrics were of little added predictive value. The algorithm was able to distinguish the MCI patients developing AD within 3 years from those patients with stable MCI over the same time-period with an area under the curve (AUC) of 0.925 and a 10-fold cross-validated accuracy of 86%, a sensitivity of 87.5%, and specificity of 85%. To our knowledge, this is the highest performance achieved so far using similar datasets. The same network provided an AUC of 1 and 100% accuracy, sensitivity, and specificity when classifying patients with AD from healthy controls. Our classification framework was also robust to the use of different co-registration templates and potentially irrelevant features/image portions.

Our approach is flexible and can in principle integrate other imaging modalities, such as PET, and diverse other sets of clinical data. The convolutional framework is potentially applicable to any 3D image dataset and gives the

* Corresponding author.
E-mail addresses: ses88@cam.ac.uk (S. Spasov), lp337@medschl.cam.ac.uk (L. Passamonti), toschi@med.uniroma2.it (N. Toschi).

[1] these authors contributed equally to this publication.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.neuroimage.2019.01.031
Received 22 August 2018; Received in revised form 9 January 2019; Accepted 12 January 2019
Available online 14 January 2019

flexibility to design a computer-aided diagnosis system targeting the prediction of several medical conditions and neuropsychiatric disorders via multi-modal imaging and tabular clinical data.

1. Introduction

More than 30 million people have a clinical diagnosis of Alzheimer's disease (AD) worldwide, and this number is expected to triple by 2050 (Barnes and Yaffe, 2011). This is due to increased life expectancy and improvements in general health care (Ferri et al., 2005). AD is a form of dementia characterized by β-amyloid peptide deposition and abnormal tau accumulation and phosphorylation which eventually lead to neuronal death and synaptic loss (Murphy and LeVine, 2010). AD-related neurodegeneration follows specific patterns which start from subcortical areas in early disease stages and spread to the cortical mantle in later stages of the disease (Braak and Braak, 1996). The classic clinical hallmark of the most common form of AD (i.e., the amnestic type) is represented by deficits in episodic memory, followed by visuo-spatial impairment, spatio-temporal orientation problems, and eventually frank dementia.

Mild cognitive impairment (MCI) is a broad, ill-defined, and highly heterogeneous phenotypic spectrum which causes relatively less noticeable memory deficits than AD. Around 10%–15% of MCI patients per year convert to AD over a relatively short time (Braak and Braak, 1995; Mitchell and Shiri-Feshki, 2008), although the annual conversion rate tends to progressively diminish. The mean conversion rate from MCI to AD is approximately 4% per year. MCI patients who do not develop AD tend to either remain stable, develop other forms of dementia, or even revert to a 'healthy' state, which suggests that MCI is a highly variable and common clinical conundrum which is likely dependent on different etio-pathogenetic mechanisms.

AD-related neuropathology can be identified several years before frank AD clinical manifestation (Braak and Braak, 1996; Delacourte et al., 1999; Morris et al., 1996; Serrano-Pozo et al., 2011; Mosconi et al., 2007), and this suggests that the development of AD might be predicted before clinical onset via *in vivo* biomarkers (e.g. PET and MR imaging as well as blood or cerebrospinal fluid (CSF) biomarkers) (Markesbery, 2010; Baldacci et al., 2018; Hampel et al., 2018; Teipel et al., 2018). Magnetic resonance imaging (MRI)-based biomarkers have attracted interest in diagnosis of AD as well in predicting MCI to AD conversion because they do not involve the use of ionizing radiation like positron emission tomography (PET), are less expensive that PET, and less invasive than the use of cerebrospinal fluid (CSF) biomarkers. MRI-based indices can also provide multi-modal information regarding the structure and function of the brain within the same scanning session, which is typically advantageous in many clinical settings.

For these reasons, there has been a growing interest in developing computational tools that are able, by using MRI-based measures, to discriminate AD patients from healthy individuals, or, most importantly, to discriminate the patients with stable MCI (sMCI) from those MCI patients who, in contrast, progress and develop AD (pMCI). To these ends, different clinical data and imaging modalities have been used so far with a variable rate of success, including for example, PET (Choi and Jin, 2018; Mosconi et al. 2004, 2007; Shaffer et al., 2013; Young et al., 2013), MRI (Filipovych and Davatzikos, 2011; Moradi et al., 2015; Mosconi et al., 2007; Tong et al., 2017, Young et al., 2013), cognitive testing (Casanova et al., 2011; Moradi et al., 2015), and CSF biomarkers (Davatzikos et al., 2011; Hansson et al., 2006; Riemenschneider et al., 2002; Sonnen et al., 2010). In this context, Moradi et al. (2015) and Tong et al., (2017) were amongst the first to: 1) perform feature selection to extract informative voxels from MRI volumes via regularized logistic regression, and 2) use the extracted voxels, along with cognitive measures, to produce support vector machine (SVM)-based predictions, achieving an area under the Receiver Operating Characteristic (ROC) curve (AUC) between 0.9 and 0.92. Similarly, Hojjati et al. (2017) employed baseline resting state functional MRI data to achieve an AUC of 0.95. In their study, the features were engineered by constructing a brain connectivity matrix which is treated as a graph, and the extracted graph measures represented the input of the SVM.

Most of these earlier studies employ a classification pipeline which relies on two independent steps. First, independent component analysis (ICA) (Shaffer et al., 2013), L1 regularization (Moradi et al., 2015; Tong et al., 2017) or morphometry (Davatzikos et al., 2011; Fan et al., 2007), is used to reduce the dimensionality of the data to a smaller set of descriptive factors. Second, these factors are fed into a multivariate pattern classification algorithm. The dimensionality reduction and classification algorithms are two separate mathematical models which involve different assumptions, and this can result in a loss of relevant information during the classification procedures (Nguyen and Torre, 2010). In addition, the most commonly employed classifiers, such as SVM (Moradi et al., 2015; Hojjati et al., 2017; Tong et al., 2017) and Gaussian Processes (Young et al., 2013), require the use of kernels, or data transformations, which are often chosen from a limited and pre-specified set. This process maps the data to a new space in which it is presumed to be easier to separate. However, constructing or choosing an application-specific kernel that acts as a reasonable similarity measure for the classification task is not always possible or easy to achieve.

The use of two separate, and methodologically disjoint, analytical pipelines as well as the need to construct *ad hoc* kernels can be avoided by employing deep learning algorithms, which have greater representational flexibility than kernel-based methods and can also automatically "learn" the necessary data transformations that maximize an arbitrary performance metric. Recently, such deep-learning methods have been applied to AD vs. healthy controls classification problems (Hosseini-Asl et al., 2016; Liu et al., 2018; Payan and Montana, 2015) and pMCI vs sMCI classification tasks (Choi and Jin, 2018; Lu et al., 2018a, b). Choi and Jin (2018) and Lu et al. (2018a) have used deep-learning to achieve one of the highest pMCI/sMCI classification performances to-date (~84%–82% conversion rate accuracies for these studies respectively). Their predictions were based on a single (albeit highly informative) imaging modality (PET). A more formal summary of the recent studies and classification methods is presented in Table 3.

The superior representational capacity of deep-learning methods typically relies on a high number of neural network parameters. Frequently, this can result in data overfitting, i.e. an apparently highly satisfactory training performance which however does not generalize well to unseen samples during testing or when applying the model. Another problem is that the data-scarce nature of medical databases is not typically sufficient to build a useful network architecture.

This study therefore aims to develop a parameter-efficient neural network architecture, based on the most recent convolutional neural network layers (i.e. the 3D separable and grouped convolutions) developed in the computer vision research field. Furthermore, we implement a dual-learning approach which simultaneously learns multi-task classification of pMCI vs. sMCI and AD vs. Health Controls (HC) by combining several input streams such as structural MRI measures as well as demographic, neuropsychological, and APOe4 genetic data (the APOe4 gene polymorphism is the only known genetic risk factor for AD in sporadic cases of AD). This new network design yields superior performance on generic visual discrimination tasks like ImageNet (Russakovsky et al., 2015; Chollet, 2017) while maintaining the number of overall network parameters low to efficiently limit the data-overfitting problem. Finally, we develop a novel feature extractor sub-network and we combine the Tensorflow (Abadi et al., 2016) and Keras (Chollet et al., 2015) libraries with our own implementation of 3D separable convolutions which is freely available at https://github.com/simeon-spasov/MCI.

Table 1
Demographic, neuropsychological and cognitive assessment as well as APOe4 genotyping data were used in this study. The data is presented in a mean ± std format. Abbreviations: APOe4 - Apolipoprotein E; CDRSB – Clinical Dementia Rating Sum of Boxes; ADAS – Alzheimer's Disease Assessment Scale; RAVLT – Ray Auditory Verbal Learning Test.

	No. of subjects	Age (years)	Male/ Female	years in education	APOe4 expression level			CDRSB	ADAS11	ADAS13	RAVLT			
					0	1	2				immediate	learning	forgetting	% forget
AD	192	75.6 ± 7	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	89.4 ± 21.2
HC	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	33.1 ± 27.7
pMCI	181	73.7 ± 7	108/73	15.9 ± 2.8	61	90	30	2 ± 1	13.5 ± 4.2	21.9 ± 5.5	27.2 ± 6.5	2.9 ± 2.2	4.9 ± 2.1	78.3 ± 27
sMCI	228	72.2 ± 7	132/96	16 ± 2.8	145	67	16	1.2 ± 0.6	8.4 ± 3.3	13.5 ± 5.3	38.5 ± 10	4.75 ± 2.5	4.35 ± 2.6	50 ± 30

2. Methods

2.1. Participants and data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The data comprised 435 men and 350 women aged between 55 and 91 years. The majority of subjects identified as white (>94%) and non-Hispanic (99.98%). All data we used is summarized in Table 1. Differences in median age across groups were tested using Friedman's ANOVA and group × gender interactions were tested using Fisher's exact test. None of these interactions resulted statistically significant ($p > 0.05$). For all participants, we employed the Magnetization Prepared Rapid Gradient-Echo (MPRAGE) T1-weighted image (structural MRI) as well as the following data: demographic data (age, gender, ethnic and racial categories, education), neuropsychological cognitive assessment tests like the dementia rating scale (CDRSB), the Alzheimer's disease assessment scale (ADAS11, ADAS13), episodic memory evaluations in the Rey Auditory Verbal Learning Test (RAVLT), as well as APOe4 genotyping. All data used in this study is from baseline assessments (no longitudinal data is used).

3. Data preprocessing

Prior to classification, all T1 weighted (T1w) images were registered to a common space (i.e. T1 template). In detail, two different T1 templates were used in order to assess the robustness of our classification methodology to coregistration inaccuracies. First, we built a custom T1 template specific to this study. To this end, we employed all T1w images, which (after N4 bias field correction) were nonlinearly co-registered to each other and averaged iteratively (i.e. the group average was recreated at the end of each iteration). The procedure was based on symmetrical diffeomorphic mapping and employed five total iterations. The second template was the Montreal Neurological T1 Template (MNI152_T1_1 mm). After the creation of both templates, all single-subject T1w images were nonlinearly registered to both templates. After co-registration to both templates we also extracted the local Jacobian Determinant (JD) images of the nonlinear part of the deformational field taking each image into template space, and masked out all non-brain areas using brainmasks generated in template space using BET, part of FSL (Jenkinson et al., 2012). The JD maps were used to complement the MRI images as an additional input stream in our model (see below). Additionally, in order to evaluate how much *a priori* knowledge about AD brain pathophsyiology could improve our classification and also how much irrelevant features hamper classification performance, we defined a set of regions of interest (ROIs) which included only brain areas known to be heavily involved in AD-related atrophy, namely parietal, temporal and frontal lobes in order to perform an inclusion test (see Fig. 4). This was based on the Hammers et al., (2003) atlas© Copyright Imperial College of Science, Technology and Medicine 2007 (www.brain-development.org).

All template creation and registration procedures were performed using the ANTs package (Avants et al., 2010, 2011). In detail, the high-dimensional non-linear transformation (symmetric diffeomorphic normalization transformation) model was initialized through a generic linear transformation which consisted of center of mass alignment, rigid, similarity and fully affine transformations followed by (metric:

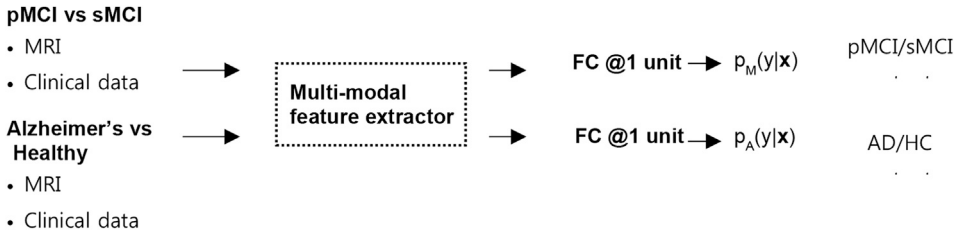

Fig. 1. Overview of our multi-tasking neural network methodology. We have designed a sub-network (the multi-modal feature extractor) to extract 4-d feature representations from the inputs of both tasks/datasets. This sub-network (with $\boldsymbol{\theta}$ network parameters) is applied on the data from both the pMCI/sMCI and AD vs healthy discrimination problems, as we assume the underlying factors of the conditions are similar, hence similar data transformations are likely to be useful. We then employ two fully connected layers, parametrized by $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$, with sigmoid outputs. The sigmoid outputs approximate the conditional distribution of the labels for the two problems given the inputs ($p_A(y|\mathbf{x})$ for the AD vs healthy task and $p_M(y|\mathbf{x})$ for the pMCI vs sMCI task). We learn the network parameters such that our model outputs correspond to the true labels in the dataset by minimizing the binary cross-entropy between the observed and estimated targets. The multi-modal feature extractor is represented by a dashed-line rectangle in Fig. 1 and Fig. 3.

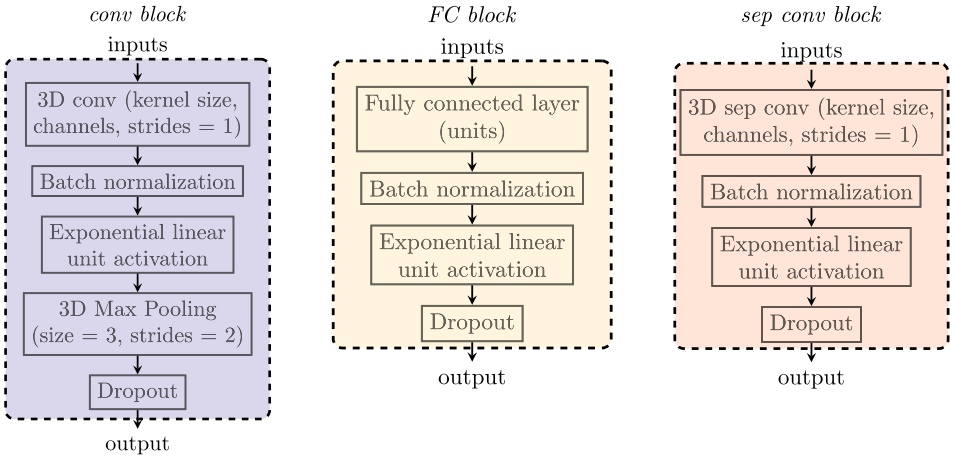

Fig. 2. Implementation of the convolutional, fully connected and separable convolutional blocks (conv block, FC and sep conv block respectively). These blocks comprise several sequential operations – firstly a (separable) convolution or dense layer followed by batch normalization and an ELU activation function. Conv blocks use 3D Max pooling with a window size of 3 and strides of 2 to gradually decrease input image dimensionality. Dropout is applied in all operational blocks. Convolutional, fully connected and max pooling layers require us to define hyperparameters, such as kernel size, number of units. These are given in brackets with some commonly used default values for our network design.

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 normalized energy profile over the last 10 iterations <10–8). Co-registration of all scans required approximately 19200 h of CPU time on a high-performance parallel computing cluster.

Numerical normalization for the co-registered MRI images was performed per sample, i.e. each 3D volume was standardized to 0 mean and unit standard deviation. The reasoning behind this is that brain atrophy could be recognized as an in-sample shift in intensity for a certain area compared to other regions. The normalization applied to the clinical features, i.e. the demographic, neuropsychological, and APOe4 genotyping data, also follows the same feature scaling procedure, where the values of each separate clinical factor are normalized between [0, 1]. On the other hand, the extracted JD images were feature-scaled to have voxel values in the [0; 1] range via subtracting the smallest value in the entire JD image set, and dividing by the difference between the largest and smallest values (also in the entire JD image set). This retains class-wise differences in volumetric changes created when co-registering an image to a template while rescaling the data to a global maximum and minimum.

4. Deep learning architecture

4.1. Architecture overview

A high-level overview of the network design is shown in Fig. 1. In this paper, we developed a feature extractor sub-network (referred to as the *multi-modal feature extractor* in Fig. 1), inspired by the parameter-efficient separable and grouped convolutional layers presented in AlexNet (Krizhevsky et al., 2012) and Xception (Chollet, 2017; Velickovic et al., 2016). In detail, the layers of the feature extractor are shared between two tasks - MCI-to-AD conversion prediction and AD/HC classification (see Figs. 2 and 3). The assumption is that both problems share common underlying factors, i.e. the MCI subjects who convert lie on a continuum between HC and AD. This means similar data transformations are likely to be useful for prediction of the two different problems. Also, this procedure increases the number of samples the extractor network is trained on, hence reducing overfitting. In addition, balancing between the two tasks can be seen as imposing soft constraints on the network parameters, and if some of the factors that explain the variations in our data are shared between the two discrimination problems, overfitting is reduced further. The feature extractor sub-network extracts 4-dimensional vectors for each of the two classification problems. These resulting latent representations are then processed by two separate fully connected layers with sigmoid activations and a binary cross-entropy loss applied at the output of each. The outputs of the fully connected layers are in the 0 to 1 range. The closer the activation is to 1, the more confident the model is that the input pattern corresponds to a diseased individual (i.e. AD or pMCI, depending on the classification task), and vice versa.

4.2. Mathematical formulation of model

We will denote the input data and labels as pairs $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}^A_1, y^A_1), \ldots, (\mathbf{x}^A_N, y^A_N), \ldots, (\mathbf{x}^M_1, y^M_1), \ldots, (\mathbf{x}^A_{N'}, y^A_{N'})\}$, where $\mathbf{x}^A_i$ is the i-th observation from the Alzheimer's and healthy subset, and $\mathbf{x}^M_j$ is the j-th observation from the pMCI vs sMCI subset. Both classification problems have corresponding class labels y^A_i and $y^M_j \in \{0, 1\}$. We refer to the empirical distributions over the AD/HC and MCI subsets as $\tilde{p}_A(\mathbf{x}, y)$ and $\tilde{p}_M(\mathbf{x}, y)$ respectively. The model log likelihoods (i.e. the conditional probabilities of the target variables, y, given the input data $\mathbf{x}$ which we model with the neural network) for the two classification problems are given by:

$$\begin{aligned} &\log p_A(y^A_i|\mathbf{x}^A_i; \theta, \varphi) = f_A(y^A; \mathbf{x}^A, \theta, \varphi) \\ &= -L_A \log p_M(y^M_j|\mathbf{x}^M_j; \theta, \psi) = f_M(y^M; \mathbf{x}^M, \theta, \psi) = -U_M \end{aligned} \quad (1)$$

The likelihood functions f_A and f_M are modelled as Bernoulli distributions, parametrized by neural network-based transformations of the input data as described in Fig. 1. The goal is to learn the network parameters such that we can approximate the *true* conditional probabilities of the labels given the inputs via the likelihood functions given by eq. (1). We use $\boldsymbol{\theta}$ to denote the parameters in the multi-modal feature extractor sub-network, and $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ to denote the weights in the final fully connected layers that output the class probabilities for the Alzheimer's vs healthy and pMCI vs sMCI tasks respectively. Learning the network parameters can be represented as:

$$\operatorname{argmin}_{(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\psi})} E_{\mathbf{x},\, y \sim \tilde{p}M(\mathbf{x},\, y)} [U_M] + \alpha E_{\mathbf{x},\, y \sim \tilde{p}A(\mathbf{x},\, y)} [L_A] \quad (2)$$

As U_M and L_A represent negative log-likelihoods, the objective function given in eq. (2) can be viewed as minimizing the weighted sum between two binary cross-entropy terms between the observed and estimated (by our network) class probabilities. Intuitively, learning the network parameters is akin to maximizing the probability of observing the labels in both datasets under the model, given the input cognitive, genetic and MRI biomarkers. We also introduced the α hyperparameter to control the trade-off between the two tasks during learning, and use $\alpha = 0.25$ in all experiments. This is a heuristic choice based on the observation that the AD/HC problem is much easier than the pMCI/sMCI problem and that the model quickly achieves high validation accuracy

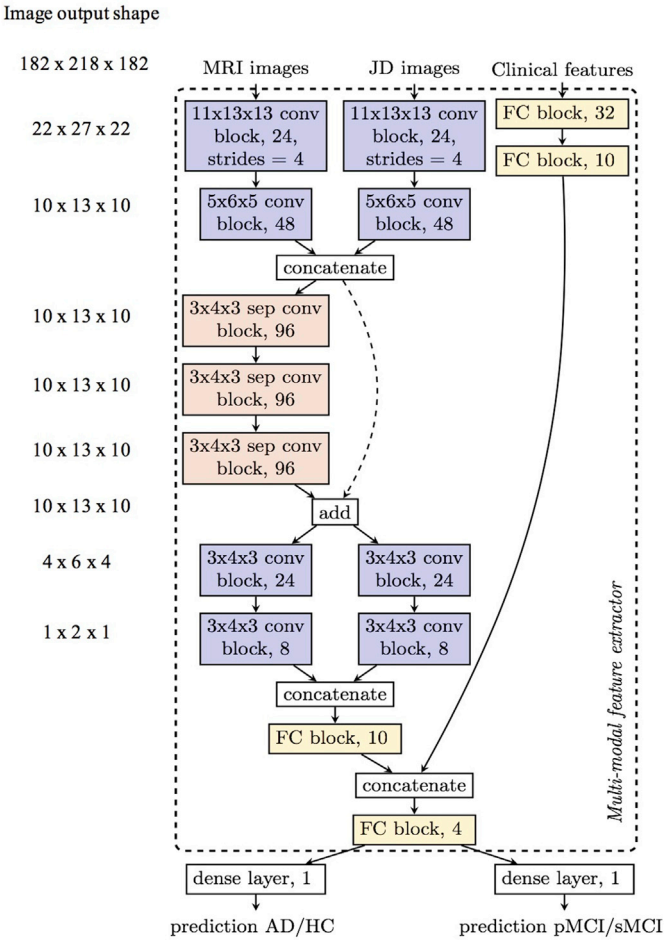

Fig. 3. Architecture of the neural network designed to take multiple 3D image volumes and tabular clinical inputs. The design of the network relies on the operational blocks shown in Fig. 2. For conv and sep conv blocks we use the notation: *kernel size, (sep) conv block, output channels.* If the strides are different from the default value of 1, the new stride value is shown in addition at the end. The *concatenation* operation works by merging the activation maps along the channel axis. Addition in the *add* block is performed element-wise between two sets of activation maps of the same size along all dimensions. The operational blocks are color-coded for the ease of the reader both in Fig. 2 and Fig. 3. Our network relies on decreasing the dimensionality of the image inputs using standard, separable and grouped convolutional blocks before concatenating the image embeddings with the clinical features compressed via fully connected blocks. The separable and grouped convolutions allow us to process the images in a parameter-efficient manner while the residual connection (dashed arrow from *concatenate* to *add*) facilitates training (Chollet, 2017). The multi-modal feature extractor sub-network (within the dashed rectangle) outputs 4-d embeddings of the input data and passes it to a dense layer which produces a prediction score. The same multi-modal feature extractor processes the inputs from both the MCI/HC and pMCI/sMCI tasks. Two different dense layers produce the final prediction scores for the two classification problems.

(see table 4 in Supplementary Material) when $\alpha = 0.25$.

4.3. 3D convolutions

Convolutional layers employed in our study work by convolving an input tensor, **x**, with a kernel of weights **W,** then adding a bias term *b,* and finally passing the result through a non-linearity. To extract a rich set of representations we repeat this process with K different kernels (also known as channels or filters) convolving the same tensor **x**, each resulting in a new *feature map* $\mathbf{h}_k$. Hence, we can write:

$$\mathbf{h}_k = f(\mathbf{W}_k * \mathbf{x} + b_k) \tag{3}$$

The feature map subscript is $k = [1, \ldots, K]$. The function f can be selected from a range of differentiable non-linear transformations, such as the sigmoid $f(u) = (1 + \exp(-u))^{-1}$ and the exponential linear unit, or ELU, (Clevert et al., 2015): $f(u) = u$ if $u \geq 0$ and $f(u) = \exp(u) - 1$ if $u < 0$. We employ the ELU transformation in our hidden layer activations and a sigmoid output for label predictions. The set of K feature maps extracted from the input **x** defines a single layer $\ell = [1, \ldots, L]$ in our convolutional neural network. Thus, the kth feature map at layer ℓ is denoted as $\mathbf{h}_k^\ell$. To construct a hierarchy of features we can use the outputs of layer ℓ-1 as inputs to layer ℓ:

$$\mathbf{h}_k^\ell = f(\mathbf{W}_k^\ell * \mathbf{h}^{\ell-1} + b_k^\ell) \tag{4}$$

where $\mathbf{h}^0$ is **x**. Note that in eq. (4), $\mathbf{h}^{\ell-1} = [\mathbf{h}_0^{\ell-1}, \ldots, \mathbf{h}_K^{\ell-1}]$ is a 4-D tensor - a collection of the K 3D feature maps extracted at layer ℓ-1. Consequently, $\mathbf{W}_k^\ell$ is also a 4-D tensor kernel of size $N^1xN^2xN^3xK$. This filter is multiplied element-wise during convolution with a $N^1xN^2xN^3$ patch in each of the K feature maps and the result is summed to produce a single scalar element (after adding a bias term and passing through a non-linear function). The convolutional procedure can be seen as sliding this kernel with strides in all three dimensions to produce $\mathbf{h}_k^\ell$. It is important to note that the number of parameters needed to extract K^ℓ feature maps in layer ℓ from the $K^{\ell-1}$ feature maps in layer ℓ-1 is given by:

$$(N^1 * N^2 * N^3 * K^{\ell-1} + 1) * K^\ell \tag{5}$$

where $N^1xN^2xN^3$ is the filter size used (see section 3.8 for actual values used in this paper).

4.4. Fully connected (dense) layers

Fully connected (FC) layers are designed to work on vectorized inputs u. Each input u_i has an associated weight w_i. In order to produce an output y_k, we form the weighted sum of all inputs $\Sigma u_i w_i$, then add a bias term b_k, and pass the result through a differentiable non-linear function like the sigmoid or the exponential linear unit. We can repeat this procedure K times with different weight parameters to produce an output vector **y**, which can be used as an input to another fully connected layer. In our work we employ these dense connections to process the tabular clinical features and to produce the final output predictions (or probability scores) of our model.

4.5. Batch normalization, dropout, L2 regularization

Several strategies are used in our network to battle overfitting. The first one is batch normalization (Ioffe and Szegedy, 2015) which normalizes a layer's outputs by subtracting their mean and dividing by their standard deviation. This whitening procedure enforces a fixed distribution of activations which stabilizes and accelerates the rate of training of deep neural nets. We also implement dropout (Srivastava et al., 2014), which works by randomly dropping units and their connections during training. An intuitive explanation of its efficacy is that each unit must learn to extract useful features on its own with different sets of randomly chosen inputs. As a result, each hidden unit is more robust to random fluctuations and learns a generally useful transformation. Finally, L2 regularization penalizes weights of high absolute value, hence directly limiting the capacity of our model, i.e. improving overfitting.

4.6. Separable convolutions

The separable convolutions we employ are similar to standard convolutional layers but reformulate the procedure in two steps by performing *depthwise* and then *pointwise* operations. Firstly, each input channel is spatially convolved separately, then the resulting outputs are mixed via *pointwise* convolutions with a kernel size of $1 \times 1 \times 1$. The depthwise procedure simply reformulates the convolutional operation

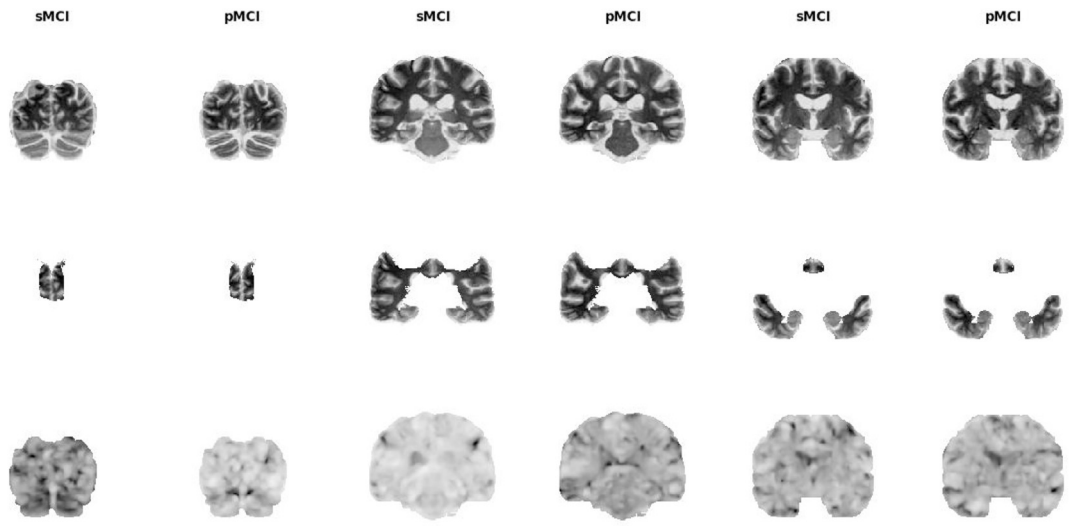

Fig. 4. Examples of the image inputs we employ in the classification framework for three different image slices. The upper row shows structural MRI images co-registered to a custom common space. The middle row displays only the brain regions we retain in the atlas-masked tests (parietal, temporal and frontal lobes). The third row shows the Jacobian Determinant images - they indicate the volumetric change a voxel in an unnormalised MRI image must undergo so as to conform to the common template.

from eq. (4) to:

$$\mathbf{h}_k^{\ell} = f\left(\mathbf{W}_k^{\ell} * \mathbf{h}_k^{\ell-1} + b_k^{\ell}\right) \tag{6}$$

Note that the difference between eq. (4) and eq. (6) is the subscript k in $\mathbf{h}_k^{\ell-1}$, denoting that feature map k in layer ℓ ($\mathbf{h}_k^{\ell}$) is only a function of feature map k in layer ℓ-1 ($\mathbf{h}_k^{\ell-1}$) in the separable convolutions case. On the other hand, standard convolutions take all $K^{\ell-1}$ feature maps as an input to produce a single output. Consequently, with our approach the parameter count in $\mathbf{W}_k^{\ell}$ is reduced to (N1*N2*N3+1)*K^{ℓ}, which is ~$K^{\ell-1}$ times more parameter-efficient as compared to standard convolutions (eq. (5)). The pointwise operation mixes all channels and requires $K^{\ell}*K^{\ell-1}$ parameters. Hence, the overall number of weights in separable convolutions is given by:

$$(N^1*N^2*N^3+1)*K^{\ell} + K^{\ell}*K^{\ell-1} \tag{7}$$

Considering the kernel sizes and number of filters in our network architecture, substituting a single conventional convolutional layer with a separable one results in ~20 times less parameters for that layer. In order to achieve the above operations, we implemented an ad-hoc 3D separable convolution module as a custom Keras layer based on a TensorFlow backend (see https://github.com/simeon-spasov/MCI).

4.7. Grouped convolutions

The grouped layer can be viewed as a compromise between standard convolutions and the separable case. This procedure splits the previous layer's feature maps in two groups (G1 and G2) along the channel axis and treats them as separate when applying further transformations (see Fig. 3). As a result, only half of the channels are used to produce a single output feature map. The grouped layer requires twice fewer parameters than the standard convolutional approach, assuming the same overall number of output feature maps is generated.

4.8. Network architecture

Since several different sequences of layers are frequently reused, they are combined in operational blocks. Each block follows a similar pattern. For instance, convolutional blocks, or conv blocks, used to processes the 3D MRI tensors, comprise a convolutional kernel with linear activations, batch normalization and an exponential linear unit (ELU) transformation with dropout. In order to reduce the resulting spatial dimensions, max pooling is used, where only the highest value in an image patch is retained, with a window of 3 pixels and a stride of 2. Each operation is applied to the outputs of the previous one. On the other hand, the clinical features undergo a series of transformations by dense, or FC (fully connected), blocks. Since these blocks act on vectorized inputs, a linear dense layer is employed instead but the same regularization precautions and activations as in the conv block are applied. We also implement a separable convolutions block, or sep conv block, which resembles the conv block but substitutes traditional convolutions with separable ones and does not rely on any pooling operations. All of these blocks are depicted in Fig. 2. Fig. 3 shows the neural network architecture we use for the AD/HC and pMCI/sMCI classification problems. Firstly, two consecutive convolutional blocks are used to reduce the dimensionality of the input MRI and Jacobian images. We then concatenate the outputs of the second conv block from the MRI and Jacobian images along the channel axis. The majority of the feature extraction is then performed by three sequential separable convolutional blocks. The dimensionality of the activation maps remains the same during this procedure. The output from the last sep conv block is summed element-wise with the activation maps from the second conv block in the add block (also known as a residual connection, introduced in He et al., 2015 and Chollet, 2017). It has been shown that residual connections facilitate training as the depth of the neural network increases. We now split the result of the summation along the channel axis in two groups to perform a grouped convolution. The motivation behind opting for grouped convolutions is to further reduce the dimensionality of the activation maps which is not possible by using the fully separable convolutions as outlined in eq. (6) but is more parameter-efficient than utilizing traditional convolutions. At this stage of the image processing pipeline the shape of the activation maps is $1 \times 2 \times 1$ with 16 channels after concatenation (8 channels in each group). We flatten the feature maps to a 32-dimensional vector and apply a fully connected block with 10 output units. This 10-dimensional vector forms the final embedding of the MRI and Jacobian images. The clinical features undergo 2 sequential transformations by fully connected blocks with 32 and 10 units respectively. The clinical features and image embeddings are concatenated and processed by a fully connected block with 4 output units. All of these operations acting on the MRI, Jacobian and clinical feature inputs which ultimately compress the input data in a 4-dimensional vector comprise the *Multi-modal feature extractor*. The parameters associated with the multi-modal feature extractor are denoted by $\boldsymbol{\theta}$ in the mathematical

formulation of our model in section 3.2. In order to obtain a prediction for each of the two tasks (AD/HC and pMCI/sMCI) we pass the 4-d output of the feature extractor sub-network through two dense (fully connected) *layers* (not blocks) with sigmoid activations and single output units. We use $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ in our mathematical formulation to denote the weights in these final fully connected layers which model the class probabilities for the AD/HC and pMCI/sMCI tasks respectively.

5. Implementation

All experiments were conducted using python version 2.7.12. The neural network was built with the Keras deep learning library using TensorFlow as backend. TensorFlow, which is developed and supported by Google, is an open-source package for numerical computation with high popularity in the deep learning community. The library allows for easy deployment on multiple graphic processing units (GPUs) (CPU-based experimentation would be prohibitive because of time constraints). The Keras wrapper provides an application programming interface (API) for quicker development and has all functionalities needed to implement the network with the exception of 3D separable convolutions, which we built as a custom layer in TensorFlow. In this paper we employed a Linux machine and two Nvidia Pascal TITAN X graphics cards with 12 GB RAM each. The model was parallelized across GPUs such that the feature extractor network works on the AD vs HC and MCI-to-AD conversion problems simultaneously to speed up training. Iterating over the whole training set once, i.e. a single epoch, takes about 30 s and prediction for a single MCI patient requires milliseconds. Since prediction would not require model parallelization or a lengthy training process, a pre-trained network is practical to be applied on a lower-end GPU (or possibly a CPU) relatively cheaply in a realistic scenario. Across all experiments certain network settings remain unchanged. These include the dropout rate - set at 0.1 for all layers and blocks; the L2 regularization penalty coefficient set at 5*10–5 for all parameters in convolutional and fully connected layers; and the convolutional kernel weight initialization which follows the procedure described by He et al., (2015). The objective function loss is minimized using the Adam optimizer by Kingma and Ba (2014) with an exponentially decaying learning rate:

$$\mathrm{lr} = 0.001{*}0.3^{\,epoch\ /\ 10} \tag{8}$$

All other parameters are kept at their default value provided in the original Adam paper (Kingma and Ba, 2014). The network hyper-parameters were picked because they resulted in sufficiently good performance on the validation set. A training batch size of 6 samples for both the AD and MCI conversion problems is randomly sampled from the dataset when training the network until the dataset is exhausted.

6. Performance evaluation

For the evaluation of the classifier, we repeated the sampling strategy to divide the samples in training, validation and test set splits. Since we have 32 samples more in the MCI dataset (16 for pMCI and 16 for sMCI) as compared to the AD/HC dataset, we used these 32 MCI subjects for testing purposes by randomly sampling 16 subjects from the pMCI and sMCI groups. The validation set comprised roughly 10% of the remaining dataset (36 subjects from MCI and AD/HC respectively) and was also generated by randomly picking in a balanced manner both from the progressive and stable MCI groups and from the healthy and AD patients as we were performing joint learning. Finally, the remaining 340 subjects from both the AD/HC and MCI subsets respectively (i.e. a total of 680 subjects) comprised the training set. No data augmentation procedures were used in this paper.

The model is trained for 40 epochs and the best performing model with the lowest objective function value (eq. (2)) on the validation set is saved and its performance is evaluated on the test set. This procedure is then repeated 10 times with different sampling seeds so as to have different samples in the train/validation/test splits (or folds) and minimize the effect of random variation. The number of subjects in each of the training/validation/testing splits in maintained the same at 680/72/32 subjects respectively. The trained model is then evaluated on the independent test set. The evaluation metrics used and reported in our results are accuracy (ACC), sensitivity (SEN), specificity (SPE). We also perform receiver operating characteristics (ROC) analysis and compute the AUC across folds. The optimal operating point of the ROC curve was found via Youden's J statistic. All accuracy, sensitivity and specificity results are reported at the optimal operating point of the ROC curve. For the AD vs HC task, we report the validation results as we only defined a test set for the pMCI/sMCI classification problem (while the AD/HC task is a helpful auxiliary problem, it turned out to be an extremely easy classification problem which is not the focus of this paper).

7. Results

Firstly, we consider the classification performance of our network on four different input biomarker combinations. The four input combinations are: 1) clinical features and T1w MRI images; 2) clinical features and Jacobian Determinant images; 3) clinical features and atlas-masked T1w MRI images; and 4) clinical features, Jacobian Determinant and T1w MRI images. We performed all of these experiments in custom template space. In order to assess the robustness of the neural network model to MRI structural misalignment, we also performed three experiments in the MNI152_T1 template space with three different input combinations (all input combinations except for number 3, i.e. clinical features and atlas-masked T1w images). Under the assumption that using the custom template will result in higher co-registration accuracy as compared to using the MNI template, the purpose of these experiments is to assess the robustness of the methodology to possible structural misalignment. In addition, we assessed the performance of our model on the AD vs healthy task with the same input variables as in the pMCI/sMCI problem. Both MNI template and AD/HC results can be found in the Supplementary Material.

7.1. Classification performance

Results are summarized in Fig. 5 and Fig. 6 and Table 2.

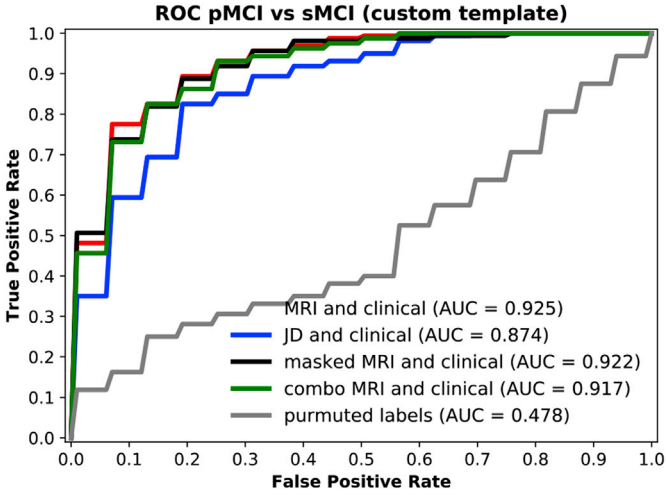

Fig. 5. ROC curves of pMCI vs sMCI classification for four input combinations: MRI images and clinical features; JD images and clinical features; Atlas-masked MRI (or just masked MRI) images and clinical features, and finally MRI and Jacobian Determinant images and clinical features. The MRI data was co-registered to our custom template prior to performing classification. The grey ROC curve at the diagonal was generated by randomly permuting the training labels for the structural MRI and clinical features input combination and predicting using this random classifier.

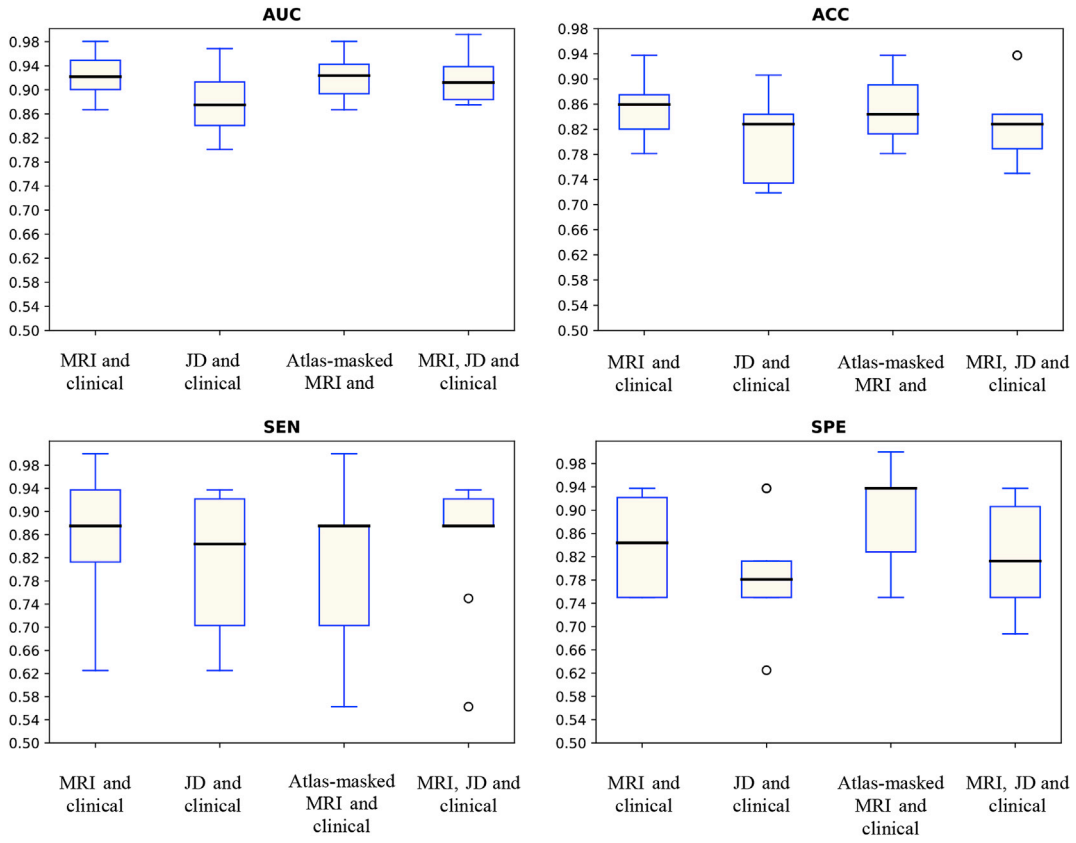

Fig. 6. Box plots for AUC, accuracy, sensitivity and specificity for pMCI vs sMCI classification based on multi-stream integration of clinical features and MRI images (co-registered to our custom template) over 10 separate test folds. The black line in each box represents the median value. The boxes encompass values between the 25th and 75th percentile whereas the tails - the top and bottom quartiles. Outliers are marked with a circle. The performance metrics correspond to the optimal operating point of each classifier.

Table 2
A comparison table between the median performance metrics on the pMCI vs sMCI classification task using our neural network model.

Input Modalities	pMCI vs sMCI							
	Custom template				MNI152 template			
	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE
MRI and clinical	0.925	86%	87.5%	84%	0.917	85%	82%	87%
Atlas-masked MRI and clinical	0.922	84%	87.5%	94%	–	–	–	–
JD and clinical	0.874	83%	84%	78%	0.881	82%	82%	81%
MRI and JD and clinical	0.917	83%	87.5%	81%	0.899	83%	77%	88%
structural MRI	0.79	72%	63%	81%				
Clinical data	0.88	81%	83%	81%				

The best performance metrics are achieved by including structural MRI along with all clinical data (demographic, neuropsychological, and APOe4 genotyping features). The median AUC across folds for the input combination comprising structural MRI images and clinical features is 0.925 whereas when we remove brain areas not classically associated with AD (i.e. using the atlas-masked images we employ in the inclusion test; see Fig. 10 in Supplementary Material), the median AUC obtained is 0.922. Comparing these results across folds using a Mann-Whitney *U* test indicated that removing brain structures unrelated to the development of AD does not hinder or aid ($P = 0.4$) discrimination in pMCI and sMCI. The median AUC when using JD images and clinical data was found to be 0.874 (Mann-Whitney test yielded p-value = 0.041 and 0.046 when compared to the input combinations comprising structural MRI and clinical data, and atlas-masked structural MRI and clinical data results, respectively). Finally, the input combination comprising all types of input streams - T1w images, JD data and clinical features resulted in an AUC of 0.917. Comparing this with the input variants comprising the structural MRI and clinical features, atlas-masked MRI and clinical features, or JD images and clinical features yielded p-values of 0.36, 0.38 and 0.07 respectively (Mann-Whitney-U test). These results suggest that adding structural MRI to the clinical features yields statistically significant higher performance as opposed to using only JD data as an image input stream. In addition, removing brain areas from structural MRI not classically associated with Alzheimer's disease did not show statistically different classification results compared to the experiments which retained all information. This suggests our model was not negatively impacted by the inclusion of irrelevant or only partially relevant features. In addition, this experiment corroborates the expectation that areas associated with AD development would possess the highest discriminative power between pMCI and sMCI, and also demonstrates a possible practical avenue for relating subsets of the input feature space to the predicted outcome with deep learning methods.

The highest median classification accuracy we achieved was 86%, which resulted from the experiments with structural MRI and clinical

Table 3
A comparative table of methodologies on the pMCI vs sMCI classification task using the ADNI dataset. We provide a performance comparison table mainly for recent studies achieving classification rates close to the state-of-the-art. The Methods column includes both the feature selection procedure(s) and the classification method.

Author	Data	AUC	ACC	SEN	SPE	Conversion time	Validation and Testing method	Method
Spasov et al. (this paper)	structural MRI + cognitive measures + APOe4 + demographics	0.925	86%	87.5%	85%	0–36 months	10-fold cross-validation	CNN
Hojjati et al. (2017)	rs-fMRI	0.95	91.4%	83.24%	90.1%	0–36 months	9-fold cross-validation (report on validation set)	Graph measures + SVM
Moradi et al. (2015)	structural MRI + cognitive measures	0.9	82%	87%	74%	0–36 months	10-fold cross-validation (report on test set)	LASSO + SVM
Liu et al. (2017)	structural MRI + FDG-PET + cognitive measures + APOe4 + demographics	0.92	84.6%	86.5%	82.4%	0–36 months	holdout	ICA + Cox model
Korolev et al. (2016)	structural MRI + clinical data + plasma-proteomic data + medications	0.87	80%	83%	76%	0–36 months	10-fold cross-validation (report on test set)	Joint Mutual Information + Kernel Learning
Beheshti et al. (2017)	structural MRI	75.08	75%	77%	73%	0–36 months	10-fold cross-validation	Morphometry + *t*-test + SVM
Choi and Jin, 2018	flurodeoxyglucose and florbetapir PET	0.89	84.2%	81%	87%	0–36 months	holdout	CNN
Tong et al. (2017)	structural MRI + cognitive measures	0.92	84%	88.7%	76.5%	0–36 months	10-fold cross-validation (report on test set)	Elastic Net + SVM
Lu et al. (2018a)	FDG-PET	–	82.5%	81.4%	83%	0–36 months	10-fold cross-validation	NN

data. The atlas-masked MRI and clinical data variant yielded the second best result with 84% classification accuracy, whereas the JD images and the clinical features gave 83% accuracy. Finally, employing all input features also resulted in an accuracy of 83%. Across the classification results from our four different input combinations the median sensitivity varies between 85% and 87.5%, and the median specificity between 78% and 94% (evaluated at the optimal operating point of each ROC curve across the test folds).

Results from the classification performance on both the custom and the MNI152 template are summarized in Table 2. We performed Mann Whitney U tests across folds on the obtained AUCs corresponding to the different input combination pairs (custom template vs MNI template).

The obtained p-values are 0.28, 0.42 and 0.24 for the structural MRI and clinical features, Jacobian Determinants and clinical features, and the combined inputs respectively. Consequently, no statistically significant difference can be found between the performance of our classifier while operating in the two different normalization spaces (custom template vs. MNI).

In addition, in order to identify the relative contribution of the structural MRI images compared to the clinical variables, we ran the pMCI vs sMCI performance evaluation procedure using either structural MRI images (not Jacobian determinant images), or clinical features as inputs. Using the clinical features alone resulted in an AUC of 0.88 (average across folds), whereas using only the structural MRI data resulted in an averaged AUC across the folds of 0.79. The ROC curves associated with these experiments can be found in Fig. 9 in the Supplementary Material.

Owing to the simpler nature of AD vs HC discrimination, regardless of the input streams and the co-registration template, results are close to 100% on all performance metrics (summarized in table 4 in Supplementary Material).

7.2. Classification variance and overfitting

Although we achieve high median performance on all metrics and on both registration templates, dispersion can be further reduced. Fig. 7 shows the standard deviation of the mean training and validation losses across the 10 test folds of the model utilizing structural MRI and clinical features as inputs, which also achieved the highest classification accuracy. It can be seen there is high overlap in the standard deviation between the training and validation losses, indicating comparable performance during both training and validation, hence no significant overfitting.

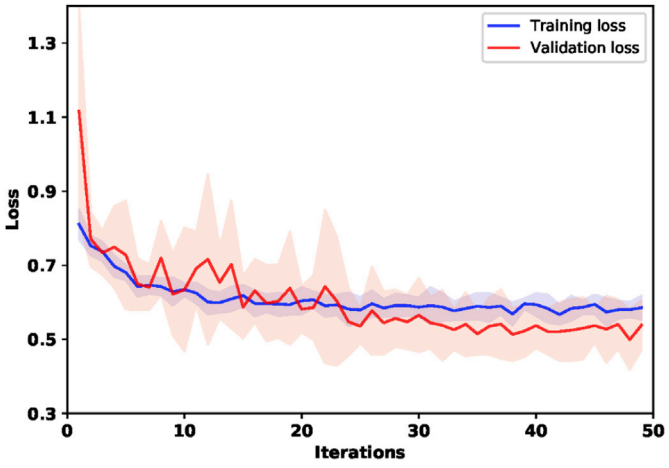

Fig. 7. Training and validation losses for our CNN architecture which utilizes structural MRI and clinical features. The standard deviation of the validation loss encompasses the red area in the image, whereas the deviation of the training loss is depicted in blue. The solid lines indicate the means of the losses across the folds.

Still, one factor which contributes to the higher validation variance compared to the training loss curve is the number of samples. Since both the validation and test sets comprise an order of magnitude less subjects than the training set, we also expect the network to manifest higher variance when evaluated on them. Secondly, although the weights were optimized using a variant of stochastic gradient descent, the hyper parameters, such as the dropout rate, the L2 regularization hyper parameter, the initial learning rate and learning rate decay were set to predefined values which gave good performance on only one of the validation folds. This was done for two reasons: 1) performing hyper parameter search at each fold was deemed prohibitive given the number of experiments we performed, and 2) hyper parameter search at each fold

may yield less clinically relevant results since this cannot be replicated in an applied clinical setting, which would require a pre-determined set of hyperparameters. As the dataset is relatively small, we observed some low level of overfitting or bias, depending on the specific data split employed. High performance metric variance is most prevalent in the sensitivity and specificity box plots since they are calculated only using either the true positives or true negatives, i.e. half the test set. Accordingly, some studies (Moradi et al., 2015; Hojjati et al., 2017; Tong et al., 2017) repeat their cross-validation loops many times (such as 100 or 1000 times) in order to further reduce their performance variance, which was not computationally feasible for our deep learning framework.

8. Discussion

Deep-learning algorithms extract a hierarchy of features from the input data via flexible and non-linear transformations. These new data representations are learnt in a manner that maximizes an arbitrary performance metric, for example binary cross-entropy. Hence, instead of relying on *a priori* knowledge or dimensionality reduction algorithms which might result in non-optimal feature selection, deep-learning uses the gradient in the performance metric to directly guide feature extraction, which can significantly improve classification results. In addition, given that the feature representations are built in a multi-layered fashion (where higher level features are derived from lower level ones), complex and information-rich data as MRI images can be dealt with and incorporated easily into the classification process.

In this paper, we have developed a new method with the primary goal to early identify the MCI patients with high risk of converting to Alzheimer's disease (AD) within 3 years, and the subsidiary aim to discriminate patients with AD from healthy controls. Our approach uses a parameter-efficient deep convolutional neural network framework, inspired by grouped and separable convolutions, to extract descriptive factors from structural MRI images acquired at the baseline clinical visit. Our work differs from previous ones because it takes into consideration potential data paucity in medical records which necessitates the use of design precautions that reduce the number of network parameters. This in turn increases the generalization capability of our model to unseen test samples (i.e. it reduces overfitting), and achieves state-of-the-art classification performance when predicting MCI-to-AD conversion. The structural MRI images were complemented by standard cognitive tests (CDRSB, ADAS, RAVLT), demographic information (age, gender, ethnicity, and education) and APOe4 genetic status data collected at the baseline visit to compute a combined score that is used to predict conversion from MCI to AD within 3 years since the baseline visit. We specifically selected these MRI and clinical measures to create a classification approach that uses the least invasive, least expensive and more commonly available diagnostic tools in the clinical practice. In other words, the MRI and clinical measures that we included here can be typically collected in non-tertiary or highly specialized medical centers, which maximizes the potential applicability of our methods in the clinical practice. For example, we did not include PET and CSF biomarkers as input measures as these measures are expensive, less diffuse, and potentially more invasive diagnostic tools than the MRI and clinical indices employed here.

We also exploited the AD and HC data to limit the effects of data overfitting. This was achieved by multi-task learning in which the same network layers are simultaneously used to extract representations from the input biomarkers for both the MCI-to-AD conversion task and the AD/HC classification problem. While previous methods employ pre-training (Payan and Montana, 2015; Hosseini-Asl et al., 2016; Liu et al., 2018) to reap similar benefits, this requires training the model twice, whereas dual-learning is a single-stage procedure which facilitate training. Furthermore, we assessed the performance of our methods using various input combinations of structural MRI, the local Jacobian Determinant of the deformational field computed during MRI co-registration, as well as the clinical data. The best results were a mean AUC of 0.925 averaged across 10 different testing folds with a mean MCI-to-AD conversion prediction accuracy of 86%, sensitivity of 87.5% and specificity of 85% (see Table 2). In addition, the use of a custom template or MNI152 template does not impact the classification results (see Table 2) which demonstrates the robustness of our network to possible structural misalignment in the MNI space. By performing masking experiments which occluded areas not typically associated with Alzheimer's Disease (refer to Fig. 10 in Supplementary Material) we also demonstrate that our framework is not negatively affected by potential inclusion of irrelevant features. This supports the idea that deep learning methods are able to identify/weight the most relevant features without being 'confounded/thrown off' by potentially misleading information from the input MRI images.

Our algorithm is also innovative in: 1) the use of parameter-efficient layers, such as grouped and separable convolutions (implemented as custom Keras layers for 3D inputs), which reduces the number of network parameters, hence overfitting; 2) the substitution of previously used network pre-training (Payan and Montana, 2015; Hosseini-Asl et al., 2016), with multi-task learning that utilizes AD/HC data to converge at a single-stage training approach, and 3) the utilization of the Jacobian Determinant as a complementary imaging input stream to maximize the extracted information from the structural MRI.

Intuitively, neural network-based methods should perform better than conventional approaches for feature extraction followed by a separate classifier, as the feature selection process is directly driven by the performance optimization procedures. However, this comes at the cost of a relatively high number of network parameters compared to the number of samples. As there are no formal estimates of the number of training samples required for any given convolutional architecture to achieve good generalization performances, we are driven by the meta-heuristic necessity of minimizing the number of network weights and maximizing the effective number of training examples to improve generalization on an independent test set and consequently enable applicability to clinical settings. Hence, our 3D model comprises 557,000 parameters, which is orders of magnitude lower than conventional 3D CNNs and even lower than recent 2D CNNs, such as AlexNet (Krizhevsky et al., 2012) and Xception (Chollet, 2017). This was achieved without sacrificing network depth or structural complexity but rather it was obtained via inserting efficient convolutional layers. To facilitate the learning procedures and increase the training samples, we hypothesized that using an auxiliary task and minimizing the joint training objective of the MCI-to-AD conversion and AD/HC classification tasks would have been an effective alternative to pre-training. In other words, AD/NC discrimination in our algorithm is seen as a simpler and easier to achieve classification task than MCI to AD conversion prediction. In addition, to speed up training convergence and limit data overfitting, we worked under the assumption that similar descriptive factors would have been useful for both classification problems. All in all, given the comparable performance of the network during training and validation (Fig. 7), we are highly confident that our deep learning framework does not suffer from significant overfitting (or underfitting) issues. We also assessed the performance of our framework by randomly permuting the training labels, which resulted in an AUC of 0.48 (Fig. 5). This further corroborates the idea that the network does not suffer from overfitting problems.

In the context of computer vision research, deep-learning methodologies can also be implemented to develop clinically useful diagnostic tools which use non-co-registered, or even non-pre-processed images, with the caveat that this approach might lead to image artefacts that reduced the discriminatory performance of the algorithm. In the context of our study, this could mean learning to relate clinically irrelevant confounds with disease outcomes. As with all multicentric studies, careful and unified data collection and processing is crucial to minimize this confound.

Our classification performances were higher than that reported in previous studies except for the work by Hojjati et al. (2017) who outperformed our current results via using rs-fMRI data. At the time of writing, ADNI had made publicly available only a limited set of rs-fMRI data (18 pMCI and 62 sMCI subjects) which made it difficult to predict

how their analytical framework would have scaled to larger populations. Furthermore, the study by Hojjati et al. (2017) does not explicitly mention the use of a separate test set which limits the generalizability of their findings (results are reported on a validation set instead of a dedicated test set).

To our knowledge, the study by Liu et al., (2017) presented comparable performance (at least in some metrics) to our model, with 84.6% classification accuracy vs 86% for our work. Liu et al., (2017), however, also included FDG-PET alongside the structural MRI and other biomarkers that we have employed here which might have improved their classification performance. Moradi et al., (2015) and Tong et al., (2017) both employed very similar methodology to each other and a dataset (structural MRI and cognitive tests) similar to the one we used here. Their sensitivity metrics are comparable to our model (~87%–88% sensitivity), however they achieve lower specificity (74%–76% vs. 85%–94% specificity for our model). A possible explanation is the inclusion of APOe4 and demographic data in our framework as well as the efficacy of the neural network. Also, as is discussed in Moradi et al., (2015) the diagnostic certainty (and hence labelling) and number of ADNI subjects varies across studies, thus hampering direct comparisons. We also evaluated the classification performance of our deep learning framework either solely on structural MRI inputs or clinical features. In our model, the use of structural MRI data alone resulted in an averaged AUC of 0.79 (see table 2 or Fig. 9 in Supplementary Material), which is higher than the AUC reported in a recent study employing similar types of datasets (Beheshti et al., 2017). On the other hand, when the clinical features alone were used as an input into the deep-learning model, we obtained an averaged AUC of 0.88 and an accuracy of 81%. Employing both structural MRI images and clinical features simultaneously increases the average AUC to 0.925 and accuracy to 86%.

In summary, we have developed a deep learning-based method for the prediction of MCI-to-AD conversion within 3 years, by combining baseline (i.e., obtained during the first visit) structural MRI, demographic, neuropsychological, and APOe4 genetic data from the ADNI database. We achieved a very high predictive performance with an average AUC of 0.925, prediction accuracy of 86%, sensitivity of 87.5% and specificity of 85%. We recommend the use of a more efficient neural network architecture (i.e., using the deep-learning framework) which typically uses fewer parameters than previous methods and therefore limits the problem of data overfitting. Our convolutional model is a generic framework that is applicable to any 3D image dataset and can be flexibly implemented to design computer-aided diagnostic systems to potentially tackle prediction and classification problems in any medical condition via multi-modal imaging measures and tabular clinical data.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The study can be split in three sub-initiatives - ADNI1, ADNI2 and ADNI GO. The initial phase known as ADNI1 included subjects between 55 and 90 years of age from approximately 50 sites from the US and Canada. ADNI2 and ADNI GO add new participants and funding to the study. The database is made available to researchers around the world and has a broad range of collaborators. The principle 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. Simeon Spasov is supported by the Engineering and Physical Sciences Research Council [EP/L015889/1]. Luca Passamonti is funded by the Medical Research Council grant (MR/P01271X/1) at the University of Cambridge, UK.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.01.031.

References

Abadi, M., et al., Nov 2016. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI).

Avants, B.B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., Gee, J.C., 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49, 2457–2466. https://doi.org/10.1016/j.neuroimage.2009.09.062.

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025.

Baldacci, F., Lista, S., O'Bryant, S.E., Ceravolo, R., Toschi, N., Hampel, H., 2018. Blood-based biomarker screening with agnostic biological definitions for an accurate diagnosis within the dimensional spectrum of neurodegenerative diseases. In: Biomarkers for Alzheimer's Disease Drug Development. Springer, New York, pp. 139–155. https://doi.org/10.1007/978-1-4939-7704-8_9.

Barnes, D.E., Yaffe, K., 2011. The projected effect of risk factor reduction on Alzheimer's disease prevalence. Lancet Neurol. 10, 819–828. https://doi.org/10.1016/s1474-4422(11)70072-2.

Beheshti, I., Demirel, H., Matsuda, H., 2017. Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. Comput. Biol. Med. 83, 109–119. https://doi.org/10.1016/j.compbiomed.2017.02.011.

Braak, H., Braak, E., 1995. Staging of alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16, 271–278. https://doi.org/10.1016/0197-4580(95)00021-6.

Braak, H., Braak, E., 1996. Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis. Acta Neuropathol. 92, 197–201. https://doi.org/10.1007/s004010050508.

Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., Espeland, M.A., 2011. High dimensional classification of structural MRI Alzheimer?s disease data based on large scale regularization. Front. Neuroinf. 5 https://doi.org/10.3389/fninf.2011.00022.

Choi, H., Jin, K.H., 2018. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav. Brain Res. 344, 103–109. https://doi.org/10.1016/j.bbr.2018.02.017.

Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr.2017.195.

Chollet, et al., 2015. Keras. available online at: : https://keras.io. (Accessed 11 August 2018).

Clevert, Djork-Arné, Unterthiner, Thomas, Hochreiter, Sepp, 2015. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). CoRR abs/1511.07289.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32, 2322. https://doi.org/10.1016/j.neurobiolaging.2010.05.023 e19-2322.e27.

Delacourte, A., David, J.P., Sergeant, N., Buee, L., Wattez, A., Vermersch, P., Ghozali, F., Fallet-Bianco, C., Pasquier, F., Lebert, F., Petit, H., Di Menza, C., 1999. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology 52. https://doi.org/10.1212/wnl.52.6.1158, 1158–1158.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imag. 26, 93–105. https://doi.org/10.1109/tmi.2006.886812.

Ferri, C.P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P.R., Rimmer, E., Scazufca, M., 2005. Global prevalence of dementia: a Delphi consensus study. Lancet 366, 2112–2117. https://doi.org/10.1016/s0140-6736(05)67889-0.

Filipovych, R., Davatzikos, C., 2011. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). Neuroimage 55, 1109–1119. https://doi.org/10.1016/j.neuroimage.2010.12.066.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19, 224–247. https://doi.org/10.1002/hbm.10123.

Hampel, H., Toschi, N., Baldacci, F., Zetterberg, H., Blennow, K., Kilimann, I., Teipel, S.J., Cavedo, E., Melo dos Santos, A., Epelbaum, S., Lamari, F., Genthon, R., Dubois, B., Floris, R., Garaci, F., Lista, S., 2018. Alzheimer's disease biomarker-guided diagnostic workflow using the added value of six combined cerebrospinal fluid candidates: aβ 1–42 , total-tau, phosphorylated-tau, NFL, neurogranin, and YKL-40. Alzheimer's Dementia 14, 492–501. https://doi.org/10.1016/j.jalz.2017.11.015.

Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K., Minthon, L., 2006. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. Lancet Neurol. 5, 228–234. https://doi.org/10.1016/s1474-4422(06)70355-6.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2015.123.

Hojjati, S.H., Ebrahimzadeh, A., Khazaee, A., Babajani-Feremi, A., 2017. Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. J. Neurosci. Methods 282, 69–80. https://doi.org/10.1016/j.jneumeth.2017.03.006.

Hosseini-Asl, E., Keynton, R., El-Baz, A., 2016. Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE. https://doi.org/10.1109/icip.2016.7532332.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. International Conference on Machine Learning (ICML) 448–456.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL. NeuroImage 62, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980 (2014).

Korolev, I.O., Symonds, L.L., Bozoki, A.C., 2016. Predicting progression from mild cognitive impairment to alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. PLoS One 11, e0138866. https://doi.org/10.1371/journal.pone.0138866.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Neural Inform. Proc. Syst., NIPS 60, 1097–1105.

Liu, K., Chen, K., Yao, L., Guo, X., 2017. Prediction of mild cognitive impairment conversion using a combination of independent component analysis and the cox model. Front. Hum. Neurosci. 11 https://doi.org/10.3389/fnhum.2017.00033.

Liu, M., Cheng, D., Wang, K., Wang, Y., 2018. Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis. Neuroinformatics. https://doi.org/10.1007/s12021-018-9370-4.

Lu, D., Popuri, K., Ding, G.W., Balachandar, R., Beg, M.F., 2018a. Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. Med. Image Anal. 46, 26–34. https://doi.org/10.1016/j.media.2018.02.002.

Lu, D., Popuri, K., Ding, G.W., Balachandar, R., Beg, M.F., 2018b. Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural MR and FDG-PET images. Sci. Rep. 8 https://doi.org/10.1038/s41598-018-22871-z.

Markesbery, William R., 2010. Neuropathologic alterations in mild cognitive impairment: a review. JAD 19, 221–228. https://doi.org/10.3233/JAD-2010-1220.

Mitchell, A.J., Shiri-Feshki, M., 2008. Temporal trends in the long term risk of progression of mild cognitive impairment: a pooled analysis. J. Neurol. Neurosurg. Psychiatr. 79, 1386–1391. https://doi.org/10.1136/jnnp.2007.142679.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002.

Morris, J.C., Storandt, M., McKeel, D.W., Rubin, E.H., Price, J.L., Grant, E.A., Berg, L., 1996. Cerebral amyloid deposition and diffuse plaques in "normal'' aging: evidence for presymptomatic and very mild Alzheimer's disease. Neurology 46, 707–719. https://doi.org/10.1212/wnl.46.3.707.

Mosconi, L., Perani, D., Sorbi, S., Herholz, K., Nacmias, B., Holthoff, V., Salmon, E., Baron, J.-C., De Cristofaro, M.T.R., Padovani, A., Borroni, B., Franceschi, M., Bracco, L., Pupi, A., 2004. MCI conversion to dementia and the APOE genotype: a prediction study with FDG-PET. Neurology 63, 2332–2340. https://doi.org/10.1212/01.wnl.0000147469.18313.3b.

Mosconi, L., Brys, M., Glodzik-Sobanska, L., De Santi, S., Rusinek, H., de Leon, M.J., 2007. Early detection of Alzheimer's disease using neuroimaging. Exp. Gerontol. 42, 129–138. https://doi.org/10.1016/j.exger.2006.05.016.

Murphy, M. Paul, LeVine, Harry, 2010. Alzheimer's disease and the amyloid-β peptide. JAD 19, 311–323. https://doi.org/10.3233/JAD-2010-1221.

Nguyen, M.H., de la Torre, F., 2010. Optimal feature selection for support vector machines. Pattern Recogn. 43, 584–591. https://doi.org/10.1016/j.patcog.2009.09.003.

Payan, Adrien, Montana, Giovanni, 2015. Predicting Alzheimer's Disease: a Neuroimaging Study with 3D Convolutional Neural Networks. ICPRAM.

Riemenschneider, M., Lautenschlager, N., Wagenpfeil, S., Diehl, J., Drzezga, A., Kurz, A., 2002. Cerebrospinal fluid tau and β-amyloid 42 proteins identify alzheimer disease in subjects with mild cognitive impairment. Arch. Neurol. 59, 1729. https://doi.org/10.1001/archneur.59.11.1729.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Serrano-Pozo, A., Frosch, M.P., Masliah, E., Hyman, B.T., 2011. Neuropathological alterations in alzheimer disease. Cold Spring Harbor Perspectives in Medicine 1. https://doi.org/10.1101/cshperspect.a006189 a006189–a006189.

Shaffer, J.L., Petrella, J.R., Sheldon, F.C., Choudhury, K.R., Calhoun, V.D., Coleman, R.E., Doraiswamy, P.M., 2013. Predicting cognitive decline in subjects at risk for alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. Radiology 266, 583–591. https://doi.org/10.1148/radiol.12120010.

Sonnen, Joshua A., Montine, Kathleen S., Quinn, Joseph F., Breitner, John C.S., Montine, Thomas J., 2010. Cerebrospinal fluid biomarkers in mild cognitive impairment and dementia. JAD 19, 301–309. https://doi.org/10.3233/JAD-2010-1236.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, Salakhutdinov, Ruslan, Bengio, Yoshua, Dropout, 2014. A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 1929–1958.

Teipel, S.J., Cavedo, E., Lista, S., Habert, M.-O., Potier, M.-C., Grothe, M.J., Epelbaum, S., Sambati, L., Gagliardi, G., Toschi, N., Greicius, M.D., Dubois, B., Hampel, H., Audrain, C., Auffret, A., Bakardjian, H., Baldacci, F., Batrancourt, B., Benakki, I., Benali, H., Bertin, H., Bertrand, A., Boukadida, L., Cacciamani, F., Causse, V., Cavedo, E., Cherif Touil, S., Chiesa, P.A., Colliot, O., Dalla Barba, G., Depaulis, M., Dos Santos, A., Dubois, B., Dubois, M., Epelbaum, S., Fontaine, B., Francisque, H., Gagliardi, G., Genin, A., Genthon, R., Glasman, P., Gombert, F., Habert, M.O., Hampel, H., Hewa, H., Houot, M., Jungalee, N., Kas, A., Kilani, M., La Corte, V., Le Roy, F., Lehericy, S., Letondor, C., Levy, M., Lista, S., Lowrey, M., Ly, J., Makiese, O., Masetti, I., Mendes, A., Metzinger, C., Michon, A., Mochel, F., Nait Arab, R., Nyasse, F., Perrin, C., Poirier, F., Poisson, C., Potier, M.C., Ratovohery, S., Revillon, M., Rojkova, K., Santos-Andrade, K., Schindler, R., Servera, M.C., Seux, L., Simon, V., Skovronsky, D., Thiebaut, M., Uspenskaya, O., Vlaincu, M., 2018. Effect of Alzheimer's disease risk and protective factors on cognitive trajectories in subjective memory complainers: an INSIGHT-preAD study. Alzheimer's Dementia. https://doi.org/10.1016/j.jalz.2018.04.004.

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., Initiative, A.D.N., 2017. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to alzheimer's disease. IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64, 155–165. https://doi.org/10.1109/tbme.2016.2549363.

Velickovic, P., Wang, D., Lane, N.D., Lio, P., 2016. X-CNN: cross-modal convolutional neural networks for sparse datasets. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI). https://doi.org/10.1109/ssci.2016.7849978.

Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. Neuroimage: Clinical. 2, 735–745. https://doi.org/10.1016/j.nicl.2013.05.004.