Transfer learning and magnetic resonance imaging techniques for the deep neural network-based diagnosis of early cognitive decline and dementia

Nitsa J Herzog[1][0000-0002-4503-615X] and George D Magoulas[2][0000-0003-1884-0772]

[1] Department of Computer Science, Birkbeck College, University of London, WC1E 7HZ, UK
[2] Birkbeck Knowledge Lab, University of London, WC1E 7HZ, UK
nitsa@dcs.bbk.ac.uk; g.magoulas@bbk.ac.uk

Abstract. Combining neuroimaging technologies and deep networks has gained considerable attention over the last few years. Instead of training deep networks from scratch, transfer learning methods have allowed retraining deep networks, which were already trained on massive data repositories, using a smaller dataset from a new application domain, and have demonstrated high performance in several application areas. In the context of a diagnosis of neurodegenerative disorders, this approach can potentially lessen the dependence of the training process on large neuroimaging datasets, and reduce the length of the training, validation, and testing process on a new dataset. To this end, the paper investigates transfer learning of deep networks, which were trained on ImageNet data, for the diagnosis of dementia. The designed networks are modifications of the AlexNet and VGG16 Convolutional Neural Networks (CNNs) and are retrained to classify Mild Cognitive Impairment (MCI), Alzheimer's disease (AD) and normal patients using Diffusion Tensor Imaging (DTI) and Magnetic Resonance Imaging (MRI) data. An empirical evaluation using DTI and MRI data from the ADNI database supports the potential of transfer learning methods in the detection of early degenerative changes in the brain. Diagnosis of AD was achieved with an accuracy of 99.75% and a 0.995 Matthews correlation coefficient (MCC) score using transfer learning of VGG models retrained on DTI scans. Early cognitive decline was predicted with an accuracy of 93.88% and an MCC equal to 0.8602 by VGG models processing MRI data. The proposed models can be used as additional tools to support a quick and efficient diagnosis of MCI, AD and other neurodegenerative disorders.

Keywords: MRI, DTI, transfer learning, dementia, deep learning.

1 Introduction

Mild cognitive impairment (MCI) belongs to the group of neurocognitive disorders characterized by minor problems with cognitive function, including memory, language, visual and spatial perception. Detecting MCI early is important since approximately 15% of the 65-year-olds with MCI develop dementia within a year, and 30%

of them develop it within 5 years. The most common course of dementia is Alzheimer's disease (AD). Neuroimaging technology is one of the key diagnostic approaches for the detection of early dementia. In this context, Magnetic Resonance Imaging (MRI) scans give detailed characteristics of the anatomical properties of the brain and cover around 50% of imaging data used for the diagnosis of brain diseases [1]. Also, Diffusion Tensor Imaging (DTI) provides the complex anatomy of the fiber tracts at the microstructural level and creates a brain-wide mapping of neuronal connections between the anatomical regions [2]. Both methods are widely used in the diagnosis of MCI and AD. Previous research has pointed out that in the early phases of the disease, white matter (WM) tract damage is happening earlier than gray matter (GM) destruction and the progression of WM atrophy exceeds the grey matter degeneration in patients with dementia [3, 4]. It has been highlighted that there is a significant correlation between WM changes and regional GM atrophy in patients with AD and this affects cognitive test performance [5]. At the same time, the correlation between GM atrophy and the damage of most WM tracts was not found in patients with the amnestic forms of MCI. In this vein, the study presented in this paper uses both imaging techniques for the early diagnosis of dementia.

In the last decade, a significant number of studies used machine learning methods for medical diagnosis [6, 7], with support vector machines (SVM), support vector regression (SVR), and random forest (RF) classifiers being among the most popular methods [6]. Advances in deep neural networks have opened a wide diagnostic opportunity in the classification and processing of medical imaging data offering additional benefits [7], and among them, Convolutional Neural Networks (CNNs) have demonstrated great potential in medical image analysis [8].

Transfer learning, which is at the core of this paper, became noticeable in medical diagnostics only in recent years. Its popularity is growing as it is a fast and highly effective approach [9]. Although deep neural networks can learn various combinations of features from coarse to fine, they typically require a lot of training data and specialized computing infrastructure to do so. Transfer learning strategies are based on thoroughly trained deep networks which are retrained using a smaller dataset from a new application domain. A popular strategy to implement transfer learning of pretrained networks is replacing the last three layers of the network’s architecture. This allows adjusting the existing network to the requirements of a new classification domain and has demonstrated in several applications at least comparable performance with models trained from scratch.

The paper explores the classification potential of popular CNN architectures, such as the AlexNet and the VGG16 networks, which have been trained on ImageNet data (www.image-net.org). Transfer learning enables quick adaptation of these computational models to new classes of medical imaging data from MRI or DTI with minimal image preprocessing. The aim is to understand how transfer learning with deep networks can be used to inform the design of DTI or MRI-based diagnostic tools for binary or multiclass classification of early mild cognitive impairment, Alzheimer’s disease and Normal (healthy) Controls (NC). This approach could offer new opportunities for quick and efficient diagnostics of different medical conditions including neurodegenerative disorders.

2 Deep learning for medical diagnosis

Compared to the classical machine learning algorithms, deep neural networks can provide an end-to-end solution, automating the image preprocessing and feature engineering stages, by considering those as part of the training process, and are able to achieve a high prediction rate of brain pathology. For example, deep learning can be used as a single classifier or in ensemble architectures for the diagnosis of brain degenerative diseases [10]. Deep networks can handle 2D and 3D data in order to distinguish between healthy and dement subjects [11]. At the same time, the advantages of deep learning models can be used for limited datasets by applying a layer-wise transfer learning approach [12] and image augmentation techniques [13]. Deep transfer learning models propose an effective way of image segmentation and can automatically classify brain scans focusing only on small brain regions [14]. Solutions proposed so far were tested only one imaging technology with very few attempts to classify Alzheimer's Disease using joint sets of MRI plus DTI data [14]. The current research complements this effort by focusing on the classification comparison of two transfer learning models. That is performed on two imaging technologies which are tested separately to determine a suitable image-algorithm combination for distinguishing between dementia stages.

2.1 Convolutional Neural Network for Image Classification

Convolutional Neural Networks, which are the center of our transfer learning scheme, are a class of multilayer neural networks that adopt the Deep Learning paradigm [8]. Input and output layers are tensors and are connected via several hidden layers of weighted nodes. Hidden layers perform important functions, data transformations, calculations, and analyses. The weights are learned and adapted by optimization algorithms. All layers are chained together. The output layer collects the processed information and generates the output, which can represent a prediction or a categorization depending on the application context.

The CNN architecture was specially designed for imaging data [15]. In that case, a two-dimensional grid of pixels typically represents each image. Each pixel value and location might be associated with numerical values depending on the black-and-white, grayscale or color images. While processing imaging data, a neural network architecture must follow the relevant application requirements. The first of these requirements is translational invariance. It means that the network layers should respond similarly to the same area regardless of where it appears in the image. The second requirement is based on the principle of "locality" when the earliest layers concentrate mainly on local regions, simple features, and abstractions. The local area representations can aggregate knowledge about the whole image.

Convolution functions

Image processing with multilayer neural networks typically requires transforming images into one-dimensional vectors. This kind of conversion impacts the relationship

between the image pixels and makes a neural classifier less effective in image processing, requiring a high number of parameters and extensive training time. In contrast, CNN can receive a tensor at the input and can learn spatial relations between pixels of the image. One can exploit the benefits of CNNs by designing significantly deeper neural architectures which, nevertheless, can learn fast complex relations from raw images. This allows the CNN to detect useful features automatically during training and develop an internal representation that classifies more complicated images than a normal multilayer neural network with sigmoid activations.

The main structural element of a CNN is the convolutional layer that operates using a convolutional function [16]. The convolution between two functions, measuring the overlap between *f* and *g* as a function of *X*, can be defined as:

$$(f * g)(X) = \int f(z)g(X - z)dz, \tag{1}$$

when one of the functions is flipped and shifted by the distance *z*.

For discrete objects such as, for instance, a set of infinite-dimensional vectors, the formula takes the following form:

$$(f * g)(i) = \sum_{a} f(a)g(i - a). \tag{2}$$

In the case of a 2D tensor, such as those used when for imaging, a corresponding sum with indices (*a*, *b*) will look as follows:

$$(f * g)(i, j) = \sum_{a} \sum_{b} f(a, b)g(i - a, j - b). \tag{3}$$

Pooling layer

Pooling layers help to save a global image representation by keeping all the advantages of the convolutional layers and other intermediate layers [16]. At the same time, the pooling procedure makes the image size significantly smaller and might alleviate the overfitting problem of the entire neural network.

The pooling operators are deterministic. They usually compute average or maximum values and are called “averaging pooling” or “max pooling” respectively. These average or maximum values are calculated at each layer location depending on the pooling function employed. The pooling layer significantly reduces the network layers’ size keeping the most significant spatial layer information, in an attempt to reduce overfitting.

ReLU activation layer

The activation layer uses differentiable operators to transform the weighted sum of the inputs received by a neuron to outputs. There are several popular activation func-

tions for deep neural networks, such as Rectified linear unit (ReLU), sigmoid, hyperbolic tangent (tanh), and Softmax [17].

The ReLU helps the neuronal network to learn fast and produce good performance [18]. For a given element *x*, the function can be expressed as the maximum of that element *x* and 0:

$$ReLU(x) = \max(x, 0). \tag{4}$$

Dropout Layer

Dropout is an efficient way to prevent the neural network from overfitting by applying a regularization technique [19]. A Dropout layer randomly sets some inputs to zero at each update of the training circle, reducing the network's capacity. All other inputs are scaled up to 1 such that the sum of all inputs remains the same.

The dropout technique can be applied to most layered neural architectures, such as models with convolutional layers, long short-term memory layers, recurrent layers, and fully connected layers. Dropout can be applied to the input layer in some situations, but it is never used with the output layer. The most advanced dropout technique specifies the probability at which parameters perform the dropout procedure. A standard threshold value for the retaining output is a probability of 0.5 for each hidden node of the layer, which means that the network retains all values above this level. The role of the dropout probability p can be explained with the following formula where each intermediate activation of the hidden node H is replaced by the random variable H':

$$H' = \begin{cases} 0 & with\ probability\ p \\ \frac{H}{1-p} & otherwise. \end{cases} \tag{5}$$

After the dropout procedure, the network weights will become larger than before. Therefore, the weights are usually scaled between zero and one before saving the model.

In summary, the CNN layers have the following functions: convolutional layers are used for feature engineering, pooling layers reduce the dimensions of the feature maps, activation layers normalize the feature maps by removing the negative values, output layers produce the classification result, the dropout layer reduces the model overfitting, and the fully connected layers compute a score of each class collected from convolutional layers.

2.2 Pretrained Convolutional Neural Network

Several popular CNN architectures have been used as base models in research projects and incorporated into modern machine learning packages. Most of them were originally introduced in the context of the ImageNet competition, launched in 2010, and won the first prize. ImageNet is the main forum for demonstrating advances in new supervised learning models in the area of computer vision. The performance of the pretrained models varies depending on the architecture and the choices of hy-

perparameters. A common transfer learning strategy consists of using pre-trained layers to construct a different network that might have similarities to the first layers. Pretrained models are available in many interesting configurations that can be grouped according to their architectural similarities. One of the first successful models is the AlexNet network. Another group of models is the Visual Geometry Group (VGG) networks, which were created from repeating blocks of structural elements that were originally introduced in the VGG model- a model built to detect geometric shapes. Another group is based on the GoogLeNet and differs from the previously mentioned architectures in the use of "Inception" blocks, which consist of parallel convolutional layers with filters of different sizes and max-pooling layer whose outputs are concatenated. Other approaches include the Network in Networks (NiN) which is based on small patch-wised convolutions, the Residual Networks (ResNet), which consist of different numbers of residual blocks and channels, and the Densely connected networks (DenseNet) that extend further the concept of residual blocks introduced in ResNet. An overview of the two models more relevant to this work is presented next.

AlexNet

AlexNet has been named after the first name of his developer, Alex Krizhevsky, who won the ImageNet competition in 2012 [20]. ImageNet competitors trained models on one million images of one thousand object categories. The lowest layers of the model are supposed to detect edges, texture, and colors, resembling the traditional image filters. The hidden layers learn a compact representation of the image whose property can be easily separated into the different data categories.

AlexNet requires an input image size of 224×224 pixels. The network was trained using an image augmentation approach, such as image clipping, flipping, and usage of color channels, which makes the model more robust, reducing overfitting. The AlexNet consists of eight layers, including five convolutional, two fully connected hidden, and one fully connected output layer. The network does not require manually designed features. All the feature detection and extraction procedures are done automatically. The first convolution filter (window) has a size of 11×11. It gives the possibility to capture rather big objects. The second and third convolutional layers have reduced filter sizes of 5×5 and 3×3, respectively.

Furthermore, the network has max-pooling layers inserted after the first, second, and fifth convolutional layers. All max-pooling layers have a window size of 3×3 and slide through the layers with a stride of 2. AlexNet uses ReLU activation functions. The network architecture is completed by two fully connected layers of 4096 output parameters (8192 parameters in total divided between dual GPUs). The model complexity is controlled by adding the dropout layer to the fully connected layer.

VGG

A repeated block structure characterizes the VGG network architecture [21]. It can be divided into two parts- integrated convolutional blocks and several fully connected

layers. Each building block includes a sequence of convolutional layers with the kernel of size 3×3 and padding of 1 pixel, a max-pooling layer of size 2×2 and a stride of 2 pixels. The original VGG network (VGG-11) has five blocks with 11 convolutional layers. The first two blocks of the network have one convolutional layer each, and the following three blocks include two convolutional layers each. The number of output channels of the first block is 64. This number doubles with each successive block and reaches 512 in the final one. Like the AlexNet, the VGG uses the ReLU activation function. Output parameters on the fully connected layer are equal to 4096, 4096 and 100 respectively. Compared to the AlexNet, the VGG-16 is computationally heavier. The VGG network has several modifications with exceeded number of convolutional layers, e.g., the VGG-16 and the VGG-19.

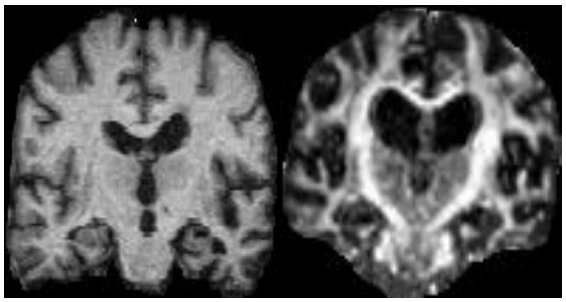
The two pretrained networks, AlexNet and VGG, are retrained and tested in the current research, as discussed in Section 4.

3 Imaging Data Repositories

Brain scans were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu)- a well-known repository of neuroimaging data. The created datasets include T1-waited images of structural MRI and DTI data of fractional anisotropy of 150 subjects. EMCI and NC patients were between 55 and 65 years old, whilst AD patients were between 65 and 90 years old. Images were processed and classified in Matlab using commodity hardware (Windows10 Enterprise, Intel (R) Core (TM), i7-7700 CPU@ 3.60 GHz, 16 GB RAM).

4 Proposed transfer learning pipeline

Initially, MRI and DTI datasets of 2D images from the ADNI3 database were created. Images were taken from the same type of 3T scanners, Siemens Medical Solutions (see details available on ADNI: http://adni.loni.usc.edu/methods/mri-tool/mri-acquisition). For the MCI and NC classes, patient data from the age group 55 to 65 years old were used to minimize the ageing effect on the imaging data. The MRI and DTI brain images were normalized using the histogram stretching technique and resized to 256×256 pixels with RGB color channels as typically done for deep learning image processing and classification. Then, the brain area in a single 2D image was segmented from the skull and other surrounding tissues using region growing and double thresholding methods (see Fig. 2). A set of 600 MRI and 600 DTI images (4 slices from each subject) were obtained from the 150 subjects. Images were balanced across classes and were used for binary and multiclass diagnosis.



(a) (b)

Fig. 2. (a) Segmented brain from MRI slice (b) Segmented brain from DTI slice.

The classification tasks were processed using transfer learning of two CNN architectures, the AlexNet and the VGG16, where the last three layers were replaced by a fully connected layer, a Softmax activation layer and an output layer, which was configured for binary or multiclass (three classes) classification depending on the type of the diagnostic task. When the cross-entropy loss function is used for training, the outputs of the Softmax layer can be interpreted as values of a probability distribution, which helps to produce the diagnostic outcome.

The weights of the pretrained AlexNet and VGG16 are used as parameters when adapting the pretrained models to the new task by retraining them using the MRI and DTI sets. Figure 3 illustrates how transfer learning of these models was used for the diagnosis of Mild Cognitive Impairment and Alzheimer's Disease; notice that the last three layers were replaced to adjust each model to the application domain.

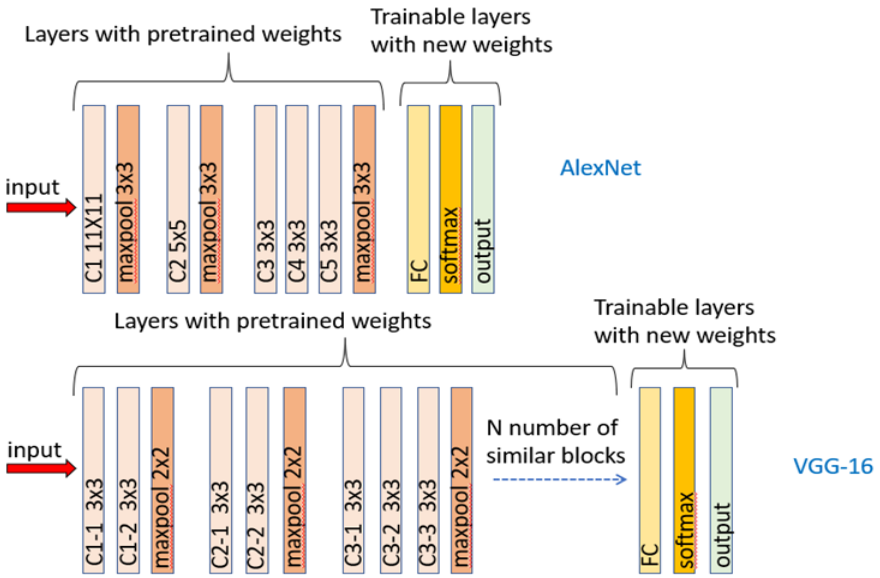


Fig. 3. Deep transfer learning architectures for diagnosis of early cognitive decline and dementia.

In general, AlexNet has been found to provide a short training time, while the VGG16 has been proved able to produce low error rates. AlexNet and VGG16 were originally configured and trained for 1000 classes using ImageNet data. AlexNet consists of 8 layers, has a size of 227MB and includes 61.0 million parameters. As mentioned in the previous section, this network requires an input image size of 227×227×3 (227 wide, 227 high, 3 color channels). The size of VGG16 is much bigger, reaching 515MB. This network has 16 layers, 138.0 million parameters, and requires an input image size of 224×224×3.

The following setup was used for retraining/finetuning both models on DTI and MRI data. All brain images were resized to fit the two pretrained networks' input sizes and were fed into the models: 80% of the images were used for training, 10% for validation, and an independent 10% of images were used for testing. Training parameters were set as follows: N = 5 for the number of training epochs for each dataset; mini-batch size = 128; validation data frequency = 50. The stochastic gradient descent with momentum (SGDM), with an initial learning rate = 0.0001, was used to train the models. All the results below are presented for the test MRI and DTI image data.

5 Experiments and Results

Experiments were conducted with the adapted configurations of AlexNet and VGG16, as described in Section 4, using DTI and MRI data. Four classification problems were tested: three binary classification tasks (EMCI vs. NC, AD vs. NC, and AD vs. EMCI) composed of 400 images each, and one multiclass task (AD vs EMCI vs NC) using 600 images with a balanced number of AD, EMCI, and NC subjects.

Table 1. Average classification performance (over 25 independent runs) on test DTI test data using transfer learning with VGG16

Model	Multiclass AD, EMCI, NC	Binary AD vs EMCI	Binary EMCI vs NC	Binary AD vs NC
Datasets	DTI	DTI	DTI	DTI
Acc	0.8438±0.020	0.7400±0.030	0.9100±0.015	0.9975±0.001
Precision	0.8600±0.023	0.7450±0.024	0.9200±0.010	1.0000±0.000
Recall	0.8329±0.030	0.7376±0.018	0.9020±0.019	0.9950±0.005
F-score	0.8462±0.017	0.7413±0.034	0.9109±0.011	0.9975±0.002
Specificity	0.7975±0.018	0.7234±0.031	0.9020±0.012	0.9950±0.003
MCC	0.6899±0.021	0.4781±0.019	0.8202±0.019	0.9950±0.005
AUROC	0.9766±0.010	0.8581±0.012	0.9700±0.010	0.9998±0.001

Table 2. Average classification performance (over 25 independent runs) on MRI test data using transfer learning with VGG16

Model	Multiclass AD, EMCI, NC	Binary AD vs EMCI	Binary EMCI vs NC	Binary AD vs NC
Datasets	MRI	MRI	MRI	MRI
Acc	0.8950±0.034	0.7813±0.015	0.9300±0.016	0.9350±0.020
Precision	0.8900±0.028	0.7950±0.025	0.9200±0.020	0.9200±0.028
Recall	0.8990±0.025	0.7737±0.027	0.9388±0.020	0.9485±0.019
F-score	0.8945±0.017	0.7842±0.019	0.9293±0.025	0.9340±0.020
Specificity	0.8641±0.019	0.7626±0.030	0.9388±0.016	0.9604±0.017
MCC	0.7922±0.023	0.5630±0.022	0.8602±0.020	0.8710±0.023
AUROC	0.9800±0.011	0.8787±0.012	0.9800±0.010	0.9756±0.013

Tables 1 and 2 summarize the models' classification performance in testing, after applying the transfer learning process described in Section 4 for the VGG network. Twenty-five independent runs were conducted by repeating the training process with a different random seed in each case. The metrics shown include accuracy rate (Acc), precision, recall, specificity, the area under the curve, which plots parametrically the true positive rate vs the false positive rate (AUC), and the F-score, which is commonly used for evaluating the performance of machine learning models. It is defined as the harmonic mean of the model's Precision and Recall (see Tables 1 and 2). Lastly, the Matthews correlation coefficient (MCC) is also reported as a metric of the quality of classifications which measures the correlation of the true classes with the predicted labels for binary classification tasks. MCC for multiclassifier was computed from averaging values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) results.

The highest performance is achieved for the binary datasets that include the AD and NC classes. Metrics for the diagnosis of EMCI vs. NC also reveal positive performance with MRI test images diagnosed more accurately than DTI test data. Distinguishing between AD and EMCI is a more challenging task, and the VGG models perform lower compared to the other binary tasks with a better performance on MRI images. Multiclassification accuracy varies from 84% for DTI data to 89.5% for MRI images. Figure 4 visualizes the classification quality as represented by the MCC for DTI and MRI data for multi- and binary classes.

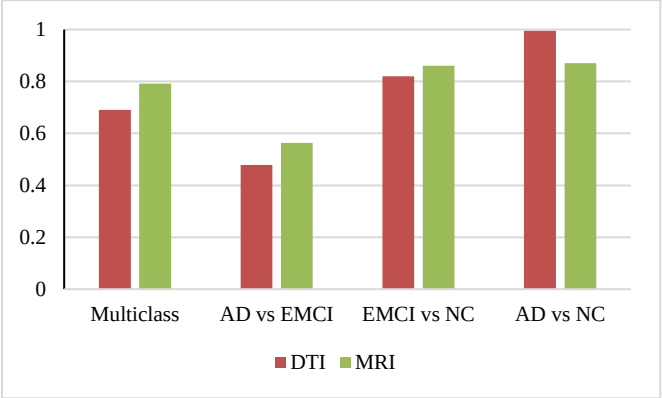


Fig. 4. VGG-based transfer learning MCC score in multiclass and binary classification of DTI and MRI data.

It is worth noticing that, as shown in Fig. 4, transfer learning with the VGG classifier detects early cognitive decline with an MCC of 0.82 using DTI images and 0.86 when MRI scans are used.

Tables 3 and 4 summarize classification performance using transfer learning with the AlexNet.

Table 3. Average classification performance (over 25 independent runs) on test DTI data using transfer learning with AlexNet

Model	Multiclass AD, EMCI, NC	Binary AD vs EMCI	Binary EMCI vs NC	Binary AD vs NC
Datasets	DTI	DTI	DTI	DTI
Acc	0.7088±0.025	0.6900±0.023	0.8500±0.025	0.9900±0.010
Precision	0.7000±0.034	0.7200±0.027	0.8700±0.019	0.9900±0.010
Recall	0.7125±0.020	0.6792±0.025	0.8365±0.025	0.9900±0.010
F-score	0.7062±0.018	0.6990±0.020	0.8529±0.250	0.9900±0.010
Specificity	0.6731±0.015	0.6800±0.023	0.8163±0.028	0.9895±0.013
MCC	0.4128±0.012	0.3801±0.018	0.7008±0.030	0.9799±0.017
AUROC	0.9052±0.018	0.8900±0.012	0.9000±0.015	0.9957±0.014

Table 4. Average classification performance (over 25 independent runs) on test MRI data using transfer learning with AlexNet

Model	Multiclass AD, EMCI, NC	Binary AD vs EMCI	Binary EMCI vs NC	Binary AD vs NC
Datasets	MRI	MRI	MRI	MRI
Acc	0.7200±0.035	0.7463±0.030	0.8500±0.018	0.8600±0.025

Precision	0.7050±0.024	0.7376±0.029	0.8600±0.015	0.8600±0.028
Recall	0.7268±0.027	0.7525±0.025	0.8431±0.016	0.8600±0.030
F-score	0.7157±0.027	0.7450±0.025	0.8515±0.019	0.8600±0.025
Specificity	0.7292±0.030	0.7738±0.023	0.8571±0.025	0.8163±0.031
MCC	0.4404±0.025	0.4943±0.020	0.7001±0.020	0.7218±0.030
AUROC	0.8550±0.016	0.8900±0.014	0.9200±0.011	0.9600±0.010

As with the VGG classifier, the classification quality between images of the AD and EMCI classes is lower compared to EMCI (EMCI vs NC) or AD (AD vs NC) class data. The classification performance for the detection of early brain changes (EMCI vs NC) can vary but without significant differences between DTI and MRI data. The diagnosis of Alzheimer's Disease using DTI data is significantly higher (99%) than using MRI data (86%) for transfer learning with the AlexNet. Comparative results in terms of MCC score in binary and multiclass classification of DTI and MRI data are shown in Figure 5.

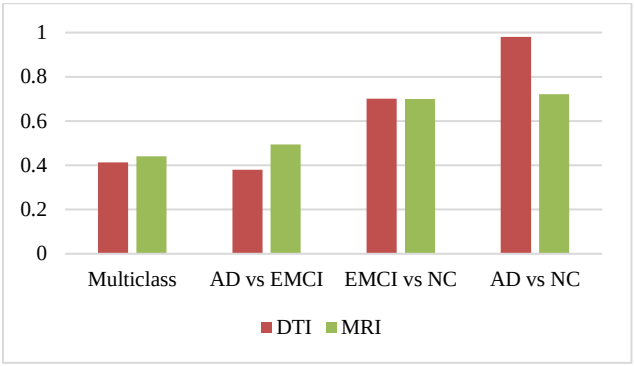


Fig.5. AlexNet-based transfer learning MCC score in multiclass and binary classification of DTI and MRI data.

Figure 5 shows that the highest MCC score for the diagnosis of AD vs NC (0.98) is achieved with DTI data, whilst almost equal detection of early brain changes (EMCI vs NC) is achieved with both types of data.

Comparing the performance of the two transfer learning models one can observe that the highest performance of 89.50 % (0.98 of AUC, 0.89 of F-score, 0.79 of MCC) in the multiclassification task is achieved with VGG-16 on MRI data. The best results in the binary classification tasks are obtained by VGG-16 nets using MRI data: AD vs EMCI (78% of accuracy, 0.88 of AUC, 0.78 of F-score); EMCI vs NC (93% of accuracy, 0.98 of AUC, 0.93 of F-score). The AD vs NC task is diagnosed better by transfer learning with the VGG-16 classifier when DTI data are used. AlexNet-based transfer learning also performs well on DTI data

It is worth noticing that the time required for training and testing differs significantly between the AlexNet-based and the VGG-based transfer learning architectures (the same commodity hardware was used for all experiments as described in Section 3). AlexNet-based transfer learning required approximately 1.3 hours for multiclassification and 0.85 hours for binary classification, whilst transfer learning with the VGG-16 took 15.7 hours and 9.7 hours respectively.

6 Discussion and Conclusions

Deep transfer learning is a promising technique for the detection of cognitive decline when MRI data are used, as demonstrated by the experimental study. Nevertheless, DTI data appear to give an advantage to deep transfer learning when used for the diagnosis of Alzheimer's disease.

The performance of the classifiers used in the research indicates the advantage of the VGG-16-based models over the AlexNet ones using the transfer learning approach. For example, MCC coefficients are 0.995 (DTI set) and 0.871 (MRI set) for the diagnosis of AD with VGG-16 in the binary case, and 0.6899 (DTI set) and 0.7922 (MRI set) in the multiclass case. Also, MCC score with retrained AlexNet are as follows: 0.9799 (DTI set) and 0.7218 (MRI set) for the diagnosis of AD, and 0.7008 (DTI set) and 0.7001 (MRI set) for the diagnosis of MCI. However, the advantage of the VGG-16 comes at a price since these models take 8 to 15 times longer to train and test than AlexNet. The exploration of the additional CNN architectures, pretrained or created from scratch, can benefit further diagnosis of cognitive decline. From a medical perspective, the findings align with previous research that showed degeneration of the white matter of the brain is connected to and correlated with gray matter atrophy in cases of Alzheimer's disease. Axons of neurons can be affected earlier than the neurons themselves and can symbolize the early onset of the disease. DTI can detect these changes quicker than MRI and become the method of choice in the early diagnosis of Alzheimer's forms of dementia. The white matter in patients with MCI is affected significantly less. Thus, in the diagnosis of MCI and the transformation of some of its forms to AD, MRI technologies help computational models perform better compared to DTI. This can be explained by the fact that cognitive decline in the case of MCI might have different morphological grounds when the destructive process does not involve white matter only. The nature of MCI is more complex and might have another, vascular reason, for amnestic and cognitive decline. Only 30% of MCI progress to AD.

Lastly, there is potential to extend this research by focusing on longitudinal studies inside image classes, based on the evaluation and analysis of the changes of WM tracts during the progression of dementia by means of transfer learning with Convolutional Neural Networks. In this context, exploring the use of additional biomarkers (features) can potentially enhance the diagnosis of MCI using Deep Learning methods.

A list of abbreviations

Alzheimer's disease (AD)
Convolutional Neural Networks (CNNs)
Densely connected networks (DenseNet)
Diffusion Tensor Imaging (DTI)
Gray matter (GM)
Matthews Correlation Coefficient (MCC)
Mild Cognitive Impairment (MCI)
Network in Networks (NiN)
Normal Controls (NC)
Random forest (RF)
Rectified linear unit (ReLU)
Residual Networks (ResNet)
Support vector machines (SVM)
Support vector regression (SVR)
Visual Geometry Group (VGG)
White matter (WM)

Acknowledgement

Data collection for this work 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). The research presented in this paper was partially funded by a BEI School Award of Birkbeck College, University of London. For the purposes of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

References

1. Segato, A., Marzullo, A., Calimeri, F., De Momi, E.: Artificial Intelligence for brain diseases: A systematic review. APL Bioengineering, vol.4, no.4, p.041503 (2020).
2. Soares, J., Marques, P., Alves, V., Sousa, N.: A hitchhiker's guide to diffusion tensor imaging. Frontiers in Neuroscience, vol.7, p.31 (2013).
3. Agosta, F., Pievani, M., Sala, S., Geroldi, C., Galluzzi, S., Frisoni, G.B., Filippi. M.: White matter damage in Alzheimer disease and its relationship to gray matter atrophy. Radiology, vol.258, no.3, pp.853-863 (2011).
4. Frings, L., Yew, B., Flanagan, E., Lam, B.Y., Hüll, M., Huppertz, H.J., Hodges, J.R., Hornberger, M.: Longitudinal grey and white matter changes in frontotemporal dementia and Alzheimer's disease. PlOS One, vol.9, no.3, p.e90814 (2014).
5. Mayo, C.D., Garcia-Barrera, M.A., Mazerolle, E.L., Ritchie, L.J., Fisk, J.D., Gawryluk, J.R. & Alzheimer's Disease Neuroimaging Initiative: Relationship between DTI metrics and cognitive function in Alzheimer's disease. Frontiers in Aging Neuroscience, vol.10, p.436 (2019).

6. Saha, A., Fadaiefard, P., Rabski, J.E., Sadeghian, A., Cusimano, M.D.: Machine learning applications using diffusion tensor imaging of human brain. A PubMed literature review. arXiv preprint arXiv:2012.10517 (2020).
7. Lundervold, A.S., Lundervold. A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, vol.29, no.2, pp.102-127 (2019).
8. Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Van Essen, B.C., Awwal, A.A.S., Asari, V.K.: The history began from alexnet: A comprehensive survey on deep learning approaches. arXiv preprint arXiv:1803.01164 (2018).
9. Marzban, E.N., Eldeib, A.M., Yassine, I.A., Kadah, Y.M. & Alzheimer's Disease Neurodegenerative Initiative: Alzheimer's disease diagnosis from diffusion tensor images using convolutional neural networks. PlOS One, vol.15, no.3, p.e0230409 (2020).
10. Nanni, L., Interlenghi, M., Brahnam, S., Salvatore, C., Papa, S., Nemni, R., ... & Alzheimer's Disease Neuroimaging Initiative: Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease. Frontiers in Neurology, 1345 (2020).
11. Yagis, E., Citi, L., Diciotti, S., Marzi, C., Atnafu, S. W., De Herrera, A. G. S.: 3D Convolutional Neural Networks for Diagnosis of Alzheimer's Disease via Structural MRI. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 65-70). IEEE (2020).
12. Mehmood, A., Yang, S., Feng, Z., Wang, M., Ahmad, A. S., Khan, R., Yaqub, M.: A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. Neuroscience, 460, 43-52 (2021).
13. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In 2018 international interdisciplinary PhD workshop (IIPhDW), pp. 117-122. IEEE (2018).
14. Aderghal, K., Afdel, K., Benois-Pineau, J., Catheline, G., & Alzheimer's Disease Neuroimaging Initiative: Improving Alzheimer's stage categorization with Convolutional Neural Network using transfer learning and different magnetic resonance imaging modalities. Heliyon, 6(12), e05652 (2020).
15. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361, no.10 (1995).
16. Zhang, A., Lipton, Z. C., Li, M., Smola, A. J.: Dive into deep learning. arXiv preprint arXiv:2106.11342 (2021).
17. Chollet, F.: Deep learning with Python. Simon and Schuster (2021).
18. Nair, V., Hinton, G. E.: Rectified linear units improve restricted Boltzmann machines. In ICML (2010).
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958 (2014).
20. Krizhevsky, A., Sutskever, I., Hinton, G. E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25 (2012).
21. Simonyan, K., & Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).