Diagnosis of Alzheimer's Disease Using Brain MR Images

Dimitrios Ioannis Belivanis* Soheil Esmaeilzadeh†

1 Introduction and Background

Alzheimer disease is one of the most growing health issues as the deaths caused by the disease almost doubled for the years 2000 to 2014 and the number of people with Alzheimer dementia is predicted to get doubled within the next twenty years in the United States [1]. However, we still lack the basic understanding for the causes and the mechanisms of the disease. Currently, diagnosis is performed mainly by behavioral observations and the person's medical history; however, brain imaging as Magnetic Resonance Imaging (MR-Image) can be used in conjunction to exclude the possibility of other brain related diseases and evaluate the progress of the disease by inspecting the evolution of the patient's brain over the years. Brain morphometric pattern analysis has been shown to have high accuracy for classification of the population with Alzheimer (AD), with mild cognitive conditions (MCI), and with normal controls (NC). The main characteristic of Alzheimer is the shrinking of brain volume caused by the loss of neurons and synapses in the cerebral cortex. The change of the brain structure can be seen with MR-Image as they are better than Computed Tomography (CT) scans for detecting the slight variations of soft tissues. With the recent improvements of machine learning we hope that we will be able to gain more insights into the problem and have earlier diagnosis in initial stages.

In previous years, there have been several approaches proposed for Alzheimer detection with Machine Learning applied on MR-Images of brain which can be classified into four different categories. First category is Voxel based methods where each individual voxel is an input for the machine learning algorithm [2]. Second group includes Region of Interest methods (ROI) where some specific regions of the brain such as the grey matter, hippocampal volume, and cortical thickness are extracted due to a priori knowledge about their effects on brain functionality and memory [3]. Third category is whole-image-level approach where the complete image is treated as a whole without considering local structures within the MR-Images [4]. The last category is patch (an intermediate scale between voxel-level and ROI-level) level approach which can be considered similar to ROI approach as parts of the image are chosen as patches based on some statistical processes [5].

Each of the four previous approaches has advantages and disadvantages. The voxel based approach in case of improper model architecture choice might overfit a training set as the feature space of the input has high dimensionality. In order to decrease the overfitting and the time for training, transfer learning approach has been used and led to good train and dev set accuracy with relatively low computational cost for the training process [6]. In ROI approach, the feature extraction is biased by the hypothesis that a certain regions of the brain are important for inspections and this can neglect the crucial information that might exist in other parts of the brain and miss extracting effective disease-related features. The whole-image approach cannot identify the subtle changes in brain structures as the appearance of brain MR images is often globally similar but locally different. In patch based approach the way to select discriminative patches from tens of thousands of patches in each MR-Image still remains a challenging problem. Moreover, most of the existing patch representations (e.g. intensity values, and/or morphological features) are based on engineered and empirically pre-defined features, which are often independent of subsequent classifier learning procedure. Furthermore, how to capture both local patch-level and global image-level features of MR-Images is also an issue with patch based approach.

For this project we are going to use a three-dimensional convolutional neural network (CNN) voxel based approach with preprocessed MR-Images which counts for all the voxels of the brain in order to capture the subtle local brain details in addition to more pronounced global details in MR-Images. By this detailed voxel based representation of MR-Images we would eliminate any a priori judgments for choosing ROIs and Patches and take into account a whole MR-Image with sub resolution of voxels. In order to avoid overfitting that might occur due to large dimension of images we will carefully design our training model's architecture in a systematic way. By doing this, we are looking forward to seeing considerable improvements in classification of the stages of the Alzheimer disease as it is reported that the early stage of AD only induces structural changes in small local regions rather than in the whole brain.

*dbelivan@stanford.edu
†soes@stanford.edu

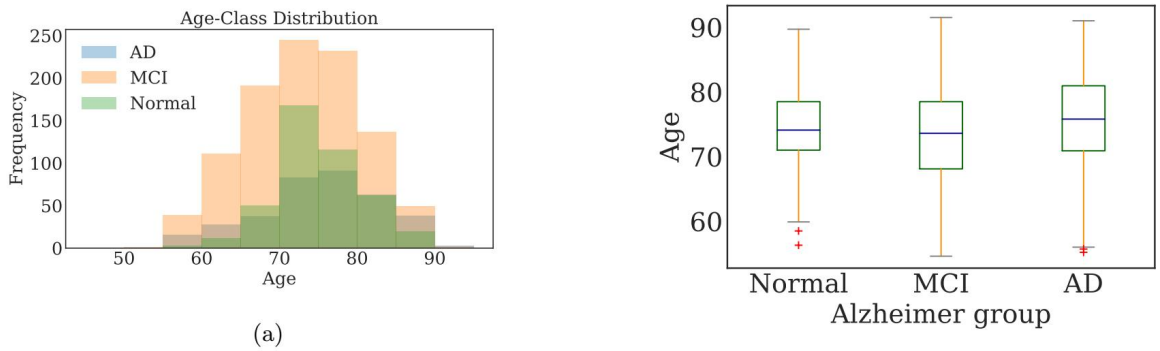
2 Dataset and Features

Two public datasets are available for this study, including the Alzheimer's Disease Neuroimaging Initiative-1 (ADNI-1) dataset [2] and the ADNI-2 dataset [3]. These two datasets contain baseline brain MR imaging from Alzheimer's disease patients and normal control subjects. We report the demographic information of studied subjects in Table 1. Subjects in the baseline ADNI-1 and ADNI-2 dataset have T1-weighted structural MR-Image data. According to some criteria (see http://adni.loni.usc.edu), such as Mini-Mental State Examination (MMSE) scores and Clinical Dementia Rating (CDR), subjects can be divided into three categories as normal conditions (NC), mild cognitive conditions (MCI), and full Alzheimer (AD). There is a total of 1797 subjects in the two dataset including 359 AD, 431 NC, and 1007 MCI subjects, which half of them are ADNI-1 that we will use in this project and postpone addition of ADNI-2 to our future work due to shortage of time, since due to its noisiness it needs extensive preprocessing before being used.

Sex	Group	Age							
		count	mean	std	min	25%	50 %	75 %	max
F	AD	165	74.3	7.9	55.2	70.7	74.5	79.5	91.0
	MCI	428	71.9	7.4	55.1	66.5	72	77.7	89.5
	NC	217	74.1	5.6	56.3	70.8	73.8	77.9	89.7
M	AD	194	76.1	7.7	56.0	71.2	77.1	82.0	90.4
	MCI	579	74.0	7.2	54.6	69.3	74.3	79.0	91.5
	NC	214	75.3	5.9	60.0	71.4	74.6	79.5	89.1

Table 1: Overview of the MR-Images data

Fig. 1a shows the number of patients in different classes of Alzheimer disease as a function of age and Fig. 1b shows the age distribution of different classes of Alzheimer disease. Almost half of the data in each male/female category is for MCI stage of the Alzheimer disease as this stage is quite difficult to be classified since it's not very similar to either of AD or NC classes of the disease. The mean age of the patients is around 75 in each classes of disease (cref. Fig. 1b) among male/female and we will include age as an extra feature in training our model. Besides, both male and female groups have approximately similar portions of patients in each of the Alzheimer classes; in spite of this we will consider gender also as a feature in training the model in addition to age and also MR-Images. Furthermore, Fig. 1a shows that although most of the patients are classified in the middle stage of the disease (i.e. MCI), a good portion of each class can be seen around the age interval of 70-80 which is a critical age range for diagnosis of Alzheimer and the number of patients visited also has its peak.

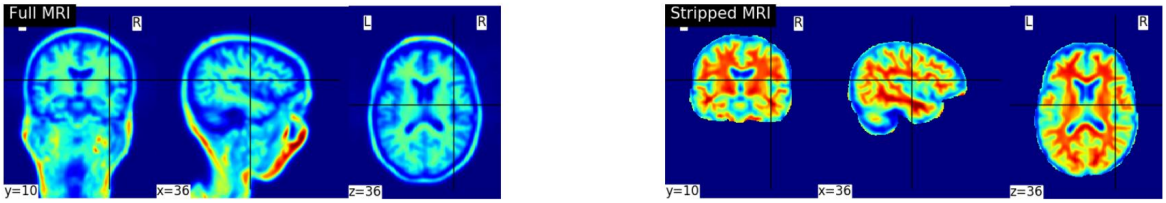


(a) (b)

Figure 1: (a) Number of patients in different classes of Alzheimer disease as a function of age, (b) age distribution of different classes of Alzheimer disease

Fig. 2 shows an illustration of an MR-Image cut in three sagittal, coronal and axial planes respectively with cut coordinates of $x = 36, y = 10, z = 36$, where in Fig. 2a a full MR-Image is shown and in Fig. 2b an skull-stripped version of MR-Image is given. Briefly, skull stripping acts as the preliminary step in numerous medical applications as it increases the speed and accuracy of diagnosis and includes removal of non-cerebral tissues like skull, scalp, and dura from brain images. Skull stripping can be part of the tissue segmentation (e.g. in SPM) but is mostly done by specialized algorithms that delineate the brain boundary. See [7] for a comparison of some brain extraction algorithms (BSE, BET, SPM, and McStrip), which suggests that all algorithms perform well in general but results highly depend on the particular dataset. In our work we use the Brain Extraction Technique (BET) proposed by Smith in 2002 [8] together with an Statistical Parametric Mapping (SPM) as a voxel based approach for brain image segmentation and extraction and choose the stripped version with higher brain tissue intensity. Performing

skull stripping on brain MR-Images reduces the size of our images by a factor of 2 and will reduce the amount of time that will be spent for training.



(a) (b)

Figure 2: (a) full and (b) skull-stripped brain MR-Image - from *left* to *right*: Coronal, Axial, and Sagittal views

3 Training Model

We have built a three-dimensional Convolutional Neural Network (3D-CNN) model in TensorFlow, two architectures are considered, (I). a complex architecture as described in Fig. (3) and (II). a simplified version which number of filters is decreased, or number of F.C. layers is decreased, or one Conv. layer at each stage is omitted.

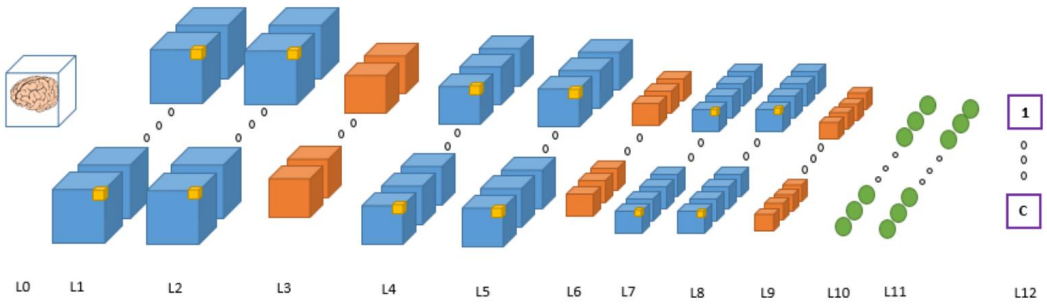


Figure 3: 3D-convolutional neural network (L_0: MR-Image (116×130×83); $L_{1,2}$: Conv. ($3^3 \times 32$); $L_{4,5}$: Conv. ($3^3 \times 64$); $L_{7,8}$: Conv. ($3^3 \times 128$); L_3, L_6, L_9: Max-pool (2^3 ,4^3 ,4^3); L_{10}, L_{11}: F.C.(512 ,128); L_{12}: Output (2))

For both architectures a *rectified linear unit (ReLU)* is used as the activation function. Our cost function of choice is cross-entropy and is minimized with the *Adam* optimizer. The hyperparameters we experiment on are the β coefficient of the L_2-regularization, the dropout probability and the size of batches, in addition to learning rate, number of filters in the convolutional layers and number of neurons in the F.C. layers.

We use F_2-score (given by Eqs. (1) and (2)) which weighs recall higher than precision (by placing more emphasis on false negatives) to evaluate the performance of our model with true positive, true negative, false positive, and false negative being as TP, TN, FP, and FN respectively.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

$$F_2 = \frac{5 \text{ precision} \times \text{recall}}{4 \text{ precision} + \text{recall}} \tag{2}$$

4 Experiments

4.1 Without Data Augmentation

Dataset has been divided into three categories of training, validation, and test set and a K-fold cross-validation process is used. Starting from the training model shown in Fig. (3) and calling it a *complex* architecture ($\mathcal{O}(10^5)$ parameters) we investigated different architectures towards simplification. First, we removed one Conv. layer from each of the three stages in Fig. (3). Furthermore, we tried different number of filters at each stage. We decreased the number of Neurons in the last two F.C. layers by half, also we removed one of the F.C. layers as well. All these attempts were toward simplifying the network (*simple* architecture with $\mathcal{O}(10^4)$ parameters) to avoid early overfitting. Besides, we investigated the effect of L_2-regularization of kernels and biases in the Conv. layers as well as the F.C. layers. First, starting with only kernel regularization with regularization coefficients of 0.01, 0.05, 0.1, 0.5, and 1.0 we regularized biases as well. Regularization coefficients 0.5 for kernels and 1.0 for the biases were found to give the best validation-set F_2-score. Afterwards, we added drop-out to the last two F.C. layers in the training process, controlling the drop-out extent by the value of keep-rate. We have tested regularized simple and complex model architectures with different keep-rate values for the F.C. layers ranging from 0.15 to 0.85 and found

that keep-rates of 0.15 and 0.25 for the first and second F.C. layers give the best validation-set accuracy in the complex model and keep-rate of 0.4 gives the best validation-set accuracy in the simple model.Worth mentioning that we add sex and age as two additional features to the last fully connected layer of the training model.

Experiments	Training-set	Validation-set
Complex	0.987	0.691
Simple	0.983	0.713
Complex + Regularization	0.981	0.751
Simple + Regularization	0.978	0.765
Complex + Drop-out + Regularization	0.967	0.821
Simple + Drop-out + Regularization	0.973	0.811

Table 2: F_2-score values of different experiments with the non-augmented dataset

4.2 With Data Augmentation

In this step, we used a data augmentation strategy where we flip the left and right hemispheres of the brain. By this augmentation of data we double the number of our labeled dataset. Table (3) shows the training set and validation set F_2-scores for both complex and simple variations of the training model in Fig. (3) which includes L_2 regularization and drop-out with the optimal values of regularization coefficients and keep-rate found in section (4.1). As it can be seen by comparing the tables (2) and (3), increasing the size of dataset by augmentation has led to improvement in the validation set F_2 score, increasing it by 12.2% from its value of 81.1% in the non-augmented case. Hence, after extensive hyperparameter tunning, and examining different training model architectures, the

Experiments	Training-set	Validation-set
Complex + Drop-out + Regularization	0.991	0.891
Simple + Drop-out + Regularization	0.998	0.933

Table 3: F_2-score values of different experiments with the augmented dataset

architecture shown in Fig. (3) with one F.C. layer and one Conv. layer in each stage with keep-rate of 0.4, regularization coefficient of 0.5 for the kernels and 1.0 for the biases has led to the highest validation-set accuracy of 0.933 for which further results will be presented in section (5).

5 Results

Fig. 4 shows the training and validations sets accuracy and loss function values with respect to the number of Epochs for the best model with validation-set accuracy of 0.933 that was discussed in section (4.2) and obtains a test-set accuracy of 0.912. Learning process is terminated when the accuracy for the training set reaches close to 1.0. Furthermore, the drop that is occurred in the loss function curve after a middle stage plateau where it reaches a saddle point can be attributed to the parameter tunning inherent to the *Adam* optimizer during a training process.

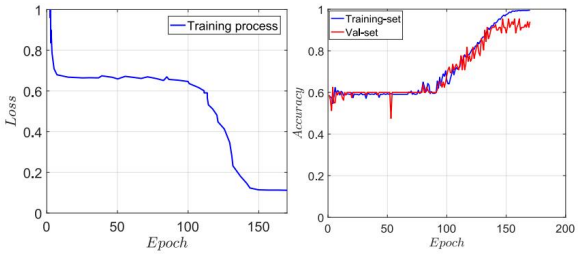


Figure 4: Training and validation sets accuracy and loss function's value with respect to the number of Epochs

The confusion matrix is provided in Fig. 5. The model is performing better in identifying the Normal Condition (NC) cases compared to identifying the Alzheimer (AD) cases but still is accurate enough in predicting AD cases with correctly predicting more than 80% of the cases.

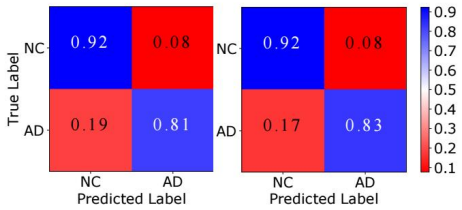


Figure 5: Normalized confusion matrix - *left*: dev-set, *right*: test-set

One of the motivation for this project is to better identify the regions of brain responsible for Alzheimer disease. For this reason we performed an image occlusion analysis on our best training model found in section (4.2) by translating a box of $5 \times 5 \times 5$ zero-valued voxels along the whole MR-Image of an AD patient that was correctly labeled as AD by the training model. White areas are irrelevant as they don't change the confidence of the prediction, red areas increase the confidence, and blue areas decrease the confidence of the model suggesting that they are areas that are important for diagnosing AD. The resulting heat map is shown in Fig. 6. As we can see the change in the confidence of the model is small suggesting that the predictions should be based on a global information from an MR-Image. Besides, it is worth noticing that the dark blue region in Fig. (6)-(a) coincide with the hippocampus part in the brain, which has been reported to be responsible for short-term memory and early stages of AD has been attributed to its malfunction [9].

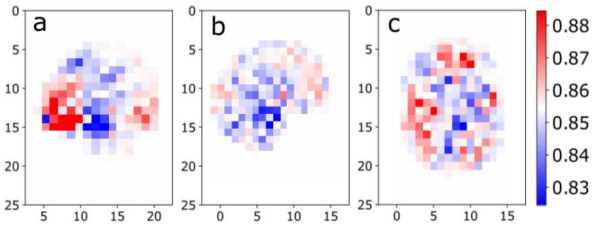


Figure 6: Probability of correct class for Alzheimer detection - (a) Coronal, (b) Axial, and (c) Sagittal views

5.1 Learning Transfer

We use the best model that is found for the binary classification of Alzheimer (i.e. AD or NC) in section (4.2) and fine-tune it using learning transfer to further subclassify the Alzheimer stage into another category called *Mild Cognitive Impairment (MCI)*, which is the stage between normal condition and full Alzheimer. Table (4) shows the validation and training set accuracy after the learning transfer on an augmented version (i.e. flipped right and left brain hemispheres) of full 3-class dataset (Fig. (1)) that includes NC, MCI, and AD. As it can be seen the simplified model architecture with drop-out and L_2-regularization gives an F_2-accuracy of 61.1% which gives a test-set accuracy of 59.12% with one F.C. layer and one Conv. layer in each stage with keep-rate of 0.25, regularization coefficient of 0.75 for the kernels and 1.0 for the biases.

Experiments	Training-set	Validation-set
Complex + Drop-out + Regularization	0.983	0.572
Simple + Drop-out + Regularization	0.991	0.611

Table 4: F_2-score values of different experiments obtained by a learning transfer for 3-class AD diagnosis with augmented dataset

6 Conclusion and Future Work

In this work, we diagnosed Alzheimer disease by MR-Images. We found critical region in brain (*hippocampus*) critical for diagnois of AD. With an extensive hyperparameter tuning and finding the best model architecture for binary classification, we fine-tuned it for further subclassifying AD into i.e. Mild Cognitive Impairment. In future, we will improving the 3-class AD diagnosis model by an extensive hyperparameter tunning which we couldn't do much due to time limits. Besides, we will request for access to ADNI2 dataset (which needs some efforts on the preprocessing as it is noisy) and further train our models with that, specially to improve the 3-class AD classifier.

7 Contributions

Dimitrios Ioannis Belivanis: Skull stripping using SPM, Data Processing and visualization, Convolutional Neural Network (CNN) building.
Soheil Esmaeilzadeh: Skull stripping using BET, Data preparation and analysis, Feature extraction and brain visualization, Training, hypter parameter tuning and visualization

8 Acknowledgements

We would like to thank Dr. Ehsan Adeli for very helpful discussions along the project, Soroosh Hemmati for his insightful comments as our mentor, and USC Mark and Mary Stevens Neuroimaging and Informatics Institute for giving us access to their dataset on Alzheimer's Disease Neuroimaging Initiative.

9 Github Link

https://github.com/soessoes/PublicCodes/tree/master/Deep%20Learning

References

[1] Alzheimer's Association. 2017 Alzheimer's Disease Facts and Figures. *Alzheimers Dement*, 13:325–373, 2017.

[2] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Landmark-based deep multi-instance learning for brain disease diagnosis. *Medical Image Analysis*, 43:157–168, 2018.

[3] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Deep multi-task multi-channel learning for joint classification and regression of brain status. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:3–11, 2017.

[4] Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Jyrki Lötjönen, and Daniel Rueckert. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. *Medical Image Analysis*, 16(4):819–830, 2012.

[5] Mingxia Liu, Jun Zhang, Ehsan Adeli, and Dinggang Shen. Deep multi-task multi-channel learning for joint classification and regression of brain status. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10435 LNCS:3–11, 2017.

[6] Ehsan Hosseini-asl, Robert Keynton, and Ayman El-baz. ALZHEIMER ' S DISEASE DIAGNOSTICS BY ADAPTATION OF 3D CONVOLUTIONAL NETWORK Electrical and Computer Engineering Department , University of Louisville , Louisville , KY , USA . (502).

[7] Kristi Boesen, Kelly Rehm, Kirt Schaper, Sarah Stoltzner, Roger Woods, Eileen Lüders, and David Rottenberg. Quantitative comparison of four brain extraction algorithms. *NeuroImage*, 22(3):1255–1261, 2004.

[8] S M Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 2002.

[9] MP Laakso, Kaarina Partanen, P Riekkinen, Maarit Lehtovirta, E-L Helkala, Merja Hallikainen, Tuomo Hanninen, Paula Vainio, and Hilkka Soininen. Hippocampal volumes in alzheimer's disease, parkinson's disease with and without dementia, and in vascular dementia an mri study. *Neurology*, 46(3):678–681, 1996.