Deep Residual Learning for Neuroimaging: An application to Predict Progression to Alzheimer's Disease

Anees Abrol[1,2], Manish Bhattarai[2], Alex Fedorov[1,2], Yuhui Du[1,3], Sergey Plis[1], Vince D. Calhoun[1,2], for the Alzheimer's Disease Neuroimaging Initiative[**]

[1]The Mind Research Network, Albuquerque, New Mexico, USA

[2]Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, New Mexico, USA

[3]School of Computer and Information Technology, Shanxi University, Taiyuan, China

Corresponding Author: Anees Abrol (*aabrol@mrn.org*)

[**]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

Contents

Abstract 3
Introduction 4
Methods 7
Structural MRI Data 7
Structural Data Pre-processing 8
Feature and Class Scores Extraction 9
Architecture Depth Selection, Regularization and Validation 11
Diagnostic/Prognostic Classification Tasks 12
Model Verification: Testing Filter Weights and Activations 12
Results 14
Architecture Depth Selection 14
Binary Diagnostic/Prognostic Classification 15
Mixed-class Prognostic Classification 16
Comparison with previous literature 17
Multi-class (4-way) Diagnostic/Prognostic Classification 18
Model Validation for the Mixed Class (modified inter-MCI) Classification Task 19
Localizing Abnormalities: Discriminative Brain Regions 21
Discussion 22
Discriminative brain regions 23
Limitations and future scope 25
Conclusion 29
ACKNOWLEDGEMENT 30
References 31
Figure Legends 37
Table 1 41
Table 2 42

ABSTRACT

This work investigates the suitability of deep residual neural networks (ResNets) for studying neuroimaging data in the specific application of predicting progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD). We focus on predicting the subset of MCI individuals that would progress to AD within three years (progressive MCI) and the subset of MCI individuals that do not progress to AD within this period (stable MCI). This prediction was conducted first as a standard binary classification task by training a ResNet architecture using MCI individuals only, followed by a modified domain transfer learning version that additionally trained on the AD and cognitively normal (CN) individuals. For this modified inter-MCI classification task, the ResNet architecture achieved better than state-of-the-art performance in predicting progression to AD using structural MRI data alone (> 7% than the second best performing method), within 1% of the state-of-the-art performance considering learning using multiple structural modalities as well, and a significant performance improvement over the classical support vector machine learning framework (p = 2.5762 x 10^{-8}). The learnt surrogate (i.e. cross-validation-fold-specific) models and the final predictive model (trained on all data) in this modified classification task showed highly similar peak activations, significant correspondence of which in the medial temporal lobe and other areas could be established with previous reports in AD literature, thus further validating our findings. Our results highlight the possibility of early identification of modifiable risk factors for understanding progression to AD using similar advanced deep learning architectures.

INTRODUCTION

Dementia is vastly underdiagnosed in most health systems mainly due to lack of educational/awareness programs and accessibility to dementia diagnostic, treatment and care services (Bradford et al. 2009; Connolly et al. 2011; Wilkins et al. 2007). Diagnosis typically occurs at relatively late stages, following which the prognosis is poor in most cases since even the state of the art (FDA-approved) medications in these stages are, at best, only modestly effective in alleviating cognitive and behavioral symptoms of the disease. As such, early therapeutic interventions can not only help improve cognitive and behavioral function of the elderly patients, but also empower them to take important decisions about their health care while they can, and significantly improve their overall quality of life.

The most widely reported form of dementia in the elderly population is Alzheimer's disease (AD) that features progressive, irreversible deterioration in memory, cognition and behavioral function. Mild cognitive impairment (MCI) has been identified as an intermediate condition between typical age-related cognitive deterioration and dementia (Markesbery 2010). This condition often leads to some form of dementia (not necessarily AD) and hence is often referred to as the prodromal stage of the disease. However, in absence of an exact (i.e. narrower) prodrome for AD, this broader population of MCI is currently an attractive target for testing preventive treatments of AD. As mentioned before, the currently approved preventive medications are effective only over a limited (early) time period (Casey, Antimisiaris, and O'Brien 2010); as such, the modest effectiveness and extremely high costs of these drugs has been a matter of constant debate especially in terms of cost to benefit balance. Hence patients showing MCI symptoms must ideally be diagnosed at early stages and be followed up on a regular basis to identify potential risks of progression to AD (or other types of dementia). Several studies are currently focused in this direction with a remarkable increase in collection and processing of multimodal neuroimaging, genetic, and clinical data. As a straightforward example, there are as many as thirty-four different live datasets that can be

accessed from the Global Alzheimer's Association Interactive Network (GAAIN) funded by the Alzheimer's Association (GAAIN Data 2017). Today, out of these splendid data collection efforts, it is especially the longitudinal studies that act as a bridge between clinical and neuropathological models (Markesbery 2010).

The structural magnetic resonance imaging (sMRI) neuroimaging modality enables tracing of brain damage (atrophy, tumors and lesions) and assists in ruling out any possible causes of dementia other than AD. This modality has additional advantages for its non-invasive nature, high spatial resolution, and ease of availability. Over the last two decades, several studies have contributed to the identification of potential AD biomarkers and prediction of progression to AD using sMRI data independently or in a multimodal pipeline (Arbabshirani et al. 2017; Falahati, Westman, and Simmons 2014; Rathore et al. 2017; Weiner et al. 2017). At the same time, the neuroimaging community has increasingly started to witness successful application of standard (i.e. classical) and advanced (i.e. deep or hierarchical) machine learning (ML) approaches to extract discriminative and diagnostic information from the high dimensional neuroimaging data (Litjens et al. 2017; Plis et al. 2014; Shen, Wu, and Suk 2017; Vieira, Pinaya, and Mechelli 2017). ML approaches are being increasingly preferred also because they allow for information extraction at the level of the individual thus making them capable of assisting the investigator in diagnostic and prognostic decision-making of the patients. The ML methods could range from standard classification frameworks (for example, logistic regression or support vector machines) that usually require manual feature engineering as a preliminary step to deep learning architectures that automatically learn optimal data representations through a series of non-linear transformations on the input data space. The last few years have seen an emergence of deep structured or hierarchical computational learning architectures to learn data representations that enable classification of brain disorders as well as predicting cognitive decline. These architectures hierarchically learn multiple levels of abstract data representations at the multiple cascaded layers, making them more suitable to learn

subtle differences in the data. Some popular deep learning architectures including multilayer perceptron, autoencoders, deep belief nets, and convolutional neural networks have indeed been applied for AD classification and predicting progression of MCI patients to AD (Chen et al. 2015; Falahati, Westman, and Simmons 2014; Li et al. 2015; M. Liu, Zhang, and Shen 2014; S. Liu et al. 2015; H. Il Suk, Lee, and Shen 2015; H. Il Suk and Shen 2013a).

Convolutional neural networks (CNNs) are a class of feed-forward artificial neural networks that have absolutely dominated the field of computer vision over the last few years with the success of strikingly superior image classification models based on models including AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ZF Net (Zeiler and Fergus 2014), VGG (Simonyan and Zisserman 2015), GoogleNet (Szegedy et al. 2015), and recently ResNet (He et al. 2016a). Deep CNN models typically stack combinations of convolutional, batch normalization, pooling and rectifier linear (ReLU) operations as a mechanism to reduce number of connections/parameters in the model while retaining the relevant invariants, and this entire network is typically followed by a fully connected layer that supports inter-node reasoning. The deep residual neural network (ResNet) learning framework as proposed by He et al., 2016a has a similar baseline architecture as the deep CNNs but additionally features parameter-free identity mappings/shortcuts that simplifies gradient flow to lower layers during the training phase. Additionally, each block of layers learns not only from the activations of the preceding block but also from the input to that preceding block. Additionally, in the original work (He et al. 2016a), these models have been shown to enable ease and simplification of neural network architecture training, thus allowing them to increase network depth and effectively enhance the overall learning performance. These networks essentially improve optimization of the "residual" mappings as compared to the collective and unreferenced original mappings (He et al. 2016a) as we will discuss next in more detail in the methods section.

Enhanced performance of the ResNet architecture within the broader imaging community motivated us to explore its diagnostic and prognostic suitability using neuroimaging data in this

work. In a systematic approach, we first comprehensively evaluate the diagnostic and prognostic performance of the ResNet architecture implemented in an open-source Pytorch GPU framework (Pytorch Resnet Architecture 2017) on a large dataset (n = 828; see Figure 1 for detailed demographics) featuring cognitively normal (CN), MCI and AD classes. Following this, we focus on prediction of progression to AD within the MCI class (i.e. predicting which MCI subjects would progress to AD within 3 years) to test the predictive performance of our learning architecture and robustness of our final predictive model (generated by fine-tuning on all available data) by comparing surrogate models (i.e. models for each cross-validation fold) with the final predictive model, and after that focus on the human brain regions maximally contributing to the prediction of MCI subjects progressing to AD as suggested by the implemented framework. Finally, we present a qualitative analysis of these results discussing the degree of success (in comparison to previously tested machine learning approaches), limitations and future scope of the evaluated framework to study the diseased brain.

METHODS

Structural MRI 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). For up-to-date information, see www.adni-info.org.

This study worked with all structural MRI scans available in the ADNI 1/2/GO/3 phases (as of November 6, 2017) that passed specific class selection criterion and the image preprocessing

pipeline quality check. Normal aging controls with no conversions in a minimum of 3 years of follow-up from their baseline scans were retained in the cognitively normal (CN) class. Subjects diagnosed as MCI with no conversions/reversions in a minimum of 3 years of follow-up from their baseline visit were grouped into the stable MCI (sMCI) class, while those converting to AD (multiple conversions excluded) within 3 years were grouped into the progressive MCI (pMCI) class. Subjects diagnosed as AD at baseline and showing no reversions in a minimum of 2 years of follow-up were retained in the AD class. Only the baseline scan for each subject was used in all analyses. Detailed scanning parameters could be accessed from ADNI data resource webpage (ADNI MRI Protocols). A total number of 830 subjects passed this criterion with further elimination of only 2 subjects that failed the image preprocessing pipeline quality analysis thus resulting in a total sample size of 828 subjects for this work. Figure 1 shows the clinical and demographic characterization of these studied CN, sMCI, pMCI and AD classes.

Structural Data Pre-processing

Image pre-processing was performed via the statistical parametric mapping 12 (SPM12) toolbox. The structural MRI images were segmented to identify the gray matter brain areas which were spatially normalized and finally smoothed using a 3D Gaussian kernel to 6 mm full width at half maximum (FWHM). The smoothed 3D gray matter images (with a voxel dimension of 160 x 195 x 170) were fed into the deep learning model for diagnostic and prognostic classification. A quality analysis correlation check was conducted with the population mean thresholded image to eliminate outlier (poorly registered) scans. This quality check discarded only 2 subjects thus retaining 828 out of the 830 subjects that satisfied the selection criterion which we use for the different diagnostic and prognostic classification tasks in this paper.

Feature and Class Scores Extraction

A non-linear, deep residual neural network (ResNet) learning framework (He et al. 2016a) was used to extract a series of relatively lower dimensional features from the very high dimensional smoothed 3D images to enhance diagnostic classification as well as identify brain regions affecting progression from MCI to AD. Similar to the deep CNNs, the deep ResNets are small blocks of multiple convolutional and batch normalization layers followed by a non-linear activation function (typically a rectified linear unit approximated by $\max(0, x)$). While traditional neural networks (NNs) learn to estimate a layer's or a small stack of layers' output activation (y) as a function (f) of the input image or activation (x) such that $y = f(x)$, ResNets, on the other hand, feature shortcut identity mappings of input space so as to enable layers to learn incrementally, or in residual representations, with the activations approximated as $y = f(x) + I(x) = f(x) + x$, where $I(*)$ is the identity function (He et al. 2016a, 2016b). As such, the latter layers in the ResNets learn not only from the output of the previous layer but also from the input to the preceding residual block, thus gaining extra information at each block in comparison to the traditional NNs. The shortcut connection approach in these networks is similar to that suggested in the "highway networks" (Srivastava, Greff, and Schmidhuber 2015), but differs in being parameter-free (i.e. shortcut connections are identity) as compared to highway networks where shortcut connections are data dependent and parameterized. It has been recently shown (Xie et al. 2017) that the aggregated transformations in this framework allow for substantially stronger representation powers in a homogenous, multi-branched architecture that strikingly requires setting a very small number of hyperparameters. We adapt this model to evaluate the architecture's performance in pair-wise (binary), mixed-class (binary but using all data by fusing more similar classes hence using more data for training) and multi-class (4-way) diagnostic classifications as shown in Table 1. While we focus on the progression of the MCI class to the AD class, all other binary classification tasks were

undertaken to confirm appropriateness of learning trends (in terms of classification performance and class separability) in the diagnostic classification of the several disease stages.

In this study, we use a modified form of an open-source Pytorch implementation of this learning framework (Pytorch Resnet Architecture 2017) evaluated for different depths, and reducing the final fully-connected layer to class probabilities to verify classification performance and appropriateness for the studied neuroimaging data. The 3D input data (smoothed gray matter maps) are fed into the deep learning ResNet framework (Figure 2) which has a series of 3D convolutional units (CUs), 3D batch-normalization units (BNUs) and non-linear activation units (Rectifier Linear Units or ReLUs) followed by a max-pooling unit (MPU) from where features are fed to the following residual blocks (RBs). Each RB has two small stacks of layers, also termed building blocks (BBs), with each BB having two CUs, two BNUs and 1 ReLU in the same specific order (CU-BNU-ReLU-CU-BNU). Following the original recommendation (Ioffe and Szegedy 2015) BNUs were adopted following every CU and before any activation functions. The activation at the output of the final residual block adder is fed into an average pooling (AP) unit for dimension reduction, and subsequently flattened (from 3D to 1D) to feed a fully connected (FC) layer featuring 512 output nodes. This relatively lower dimensional flattened feature space at the output of the first FC layer (FC1) is fed into a second FC layer (FC2) to estimate the diagnostic class probabilities/scores.

Training and testing routines were implemented on an NVIDIA CUDA parallel computing platform (accessing 3 independent servers each with 4 GeForce GTX 1080 11 GB GPUs) using GPU-accelerated CUDA toolkit/compilation and Pytorch python package tensor libraries. The Adam stochastic optimization algorithm (Kingma and Ba 2015) was preferred for its computational efficiency, relatively low memory requirements, and suitability for problems with large data/parameters size. A batch size of 16, fixed learning rate parameter of 0.001 and L2 weight decay parameter of 0.01 were chosen for the final model selection, and all further classifier

performance and feature estimation routines based on a preliminary analysis on the CN vs. AD classification task that suggested (1) insignificant effect of batch-size on learner performance, and (2) the above values of learning rate and L2 weight decay parameter through a grid-search cross-validation analysis. Due to the GPU device memory constraints, we tested only for batch sizes of 2, 4, 8 and 16 and since batch-size did not noticeably affect performance, the maximum batch-size of 16 was chosen to speed up computations (as compared to batch sizes 2, 4 and 8). Subsequently, ResNet's performance for different model depths (number of residual blocks) was compared to choose the appropriate model depth for consistent comparison across several classification tasks as demonstrated in Table 1.

Architecture Depth Selection, Regularization and Validation

The ResNet architecture with different depths (*D = 1, 2, 3, 4*; where *D* is the number of residual blocks) was tested for diagnostic classification performance for the CN vs. AD classification task. We retained the architecture depth with the best performance as suggested in this analysis *(D = 3)* for all other classification tasks for consistent comparison. Figure 2 illustrates the modular structure of the selected framework, whereas Figure 3 shows a comparison of the model performances at different depths. As shown in Figure 2, following the MPU, this architecture featured three RBs followed by two FCUs; hence, in all, thirteen convolutional and two fully connected layers were used in this fifteen layer model. Use of BNUs, default L2 weight decay (regularization) in the Adam Optimizer, repeated stratified k-fold cross-validation for the diagnostic and prognostic classification tasks and early stopping were measures undertaken to prevent any overfitting and reduce classification performance bias. This chosen architecture was then used to extract the features and class probability scores for the different binary/mixed-class/multi-class classification tasks as discussed in the following section.

Diagnostic/Prognostic Classification Tasks

Classification performance for the different binary diagnostic and prognostic classification tasks (CN vs. AD, CN vs. pMCI, sMCI vs. AD, sMCI vs. pMCI, CN vs. sMCI, and pMCI vs. AD) for the 4 studied groups was evaluated (Tasks TA1 through TA6 in Table 1). Additionally, mixed-class inter-MCI (Task TB: CN+sMCI vs. pMCI+AD; training on all CN and AD data plus 80% of sMCI and pMCI data; testing on 20% sMCI and pMCI data) and multi-class (Task TC: 4-way) classification tasks were performed to enhance classification performance and extract additional information than that conveyed by the binary classifiers respectively. Notably the mixed inter-MCI class classification task was evaluated to explore any additional benefits of domain transfer learning (Cheng et al. 2015) i.e. if training the classifier with more data samples (i.e. all CN and AD datasets) resulted in an improvement in the classification performance. While all other classification tasks were conducted to evaluate the framework performance as compared to frameworks used in similar studies in the recent literature, only the mixed/modified inter-MCI classification task that assumes the highest clinical relevance (in terms of learning about progression to AD) was focused on to seek evidence of the most affected brain areas while progressing to AD. All classification tasks were conducted using repeated (n = 10), stratified 5-fold cross-validation procedures. Classification performance metrics including accuracy, sensitivity, specificity, and balanced accuracy were computed, and additionally complemented by conducting the receiver operating characteristic (ROC) curve analysis to estimate the area under the curve (AUC) performance metric for the several undertaken classification tasks.

Model Verification: Testing Filter Weights and Activations

Following the preceding performance analysis of the several diagnostic and prognostic classification tasks, we focused only on the modified inter-MCI prognostic classification task (Table 1, Task TB) for identification of the most discriminative brain regions in progression to AD. We will refer to the models trained for each cross-validation fold as surrogate models and fine-tune the

final predictive model on all data. In this iterative learning approach, the surrogate models were additionally regularized by use of a validation-error-based early stopping mechanism to avoid overfitting and impose control over the increase in generalization error. As such, the model training phase was simply terminated at a point (iteration number) whereon there was no significant increase in the model performance (in terms of validation accuracy). The number of epochs parameter suggested by the early stopping method was averaged over the cross-validation folds, and the final predictive model was run for this averaged parameter value to allow learning of the final predictive model in a totally unsupervised way. Once the final predictive model was generated, additional tests were performed on filter weights and output activations at the first convolutional layer for each of the surrogate models as well as the final predictive model to ensure proper training which we discuss next.

During the training phase, the network iteratively learns the data and updates the filter weights in the multiple channels of convolutional layers to minimize the training error. Likewise, as in case of any CNN architecture, resembling (i.e. highly correlated or redundant) filters almost certainly indicate over-parametrization, or incorrect updating of the loss function and thus inappropriate learning directionality. Hence, we performed a similarity analysis on the weights of the 64 (size 3 x 3 x 3) filters learnt at the first convolution layer as a simple sanity check. Next, while inference on discriminative brain regions was done from the features from the final predictive models that trained on all data, projections estimated from the surrogate models must ideally be similar to confirm similar learning trends i.e. robustness of the extracted features. To inspect this, we compared the activations at the output of the first convolutional layer (post batch-normalization and non-linear activation) for the multiple surrogate models with the final predictive model. Finally, the brain regions most discriminative of likely progression to AD were localized by projecting the sub-sampled mean activations at the output of the first convolutional layer (batch-normalized and post non-linear activation) back to the brain space.

RESULTS

Architecture Depth Selection

In a repeated (n=10), stratified 5-fold cross-validation framework, the CN and AD datasets were evaluated for 100 epochs. The stratified cross-validation procedure was performed on the pooled CN and AD classes to study the effect of adding depth to the implemented architecture (i.e. further convolutional layers or residual blocks). This analysis reported significant improvement in validation accuracy by a model that used 3 residual blocks (D3: depth = 3) as compared to a model that used 2 residual blocks (D2: depth = 2; p = 1.6996e-07) and a model that used 1 residual block (D1: depth = 1; p = 4.5633e-13). Adding another residual block (i.e. depth = 4) did not result in significant improvement in performance; hence, we've settled on the D3 model and validated it in the several classification/prediction tasks, as will be shown in the forthcoming sub-sections. In this particular analysis, the models were run for 100 epochs for each depth and used the exact same training and test datasets in each of the cross-validation folds for consistency in performance comparison. A comparison of training error, training loss and validation error for the different depths is shown in Figure 3A. Additionally, the 512-dimensional feature space at the output of the first fully connected layer in the ResNet model was projected onto a two-dimensional space using the t-distributed stochastic neighbor embedding (tSNE) algorithm (der Maaten et al. 2008) to visualize class separation differences with model order. We show projections from a surrogate model (from a sample cross-validation fold) for a sample epoch around which the D3 model clearly exhibits significant differences in validation accuracy (Figure 3B); projections from other surrogate model (from other cross-validation folds) and other epochs beyond the significant difference showing epoch could be expected to exhibit a similar pattern because of evidence from results in Figure 3A.

Binary Diagnostic/Prognostic Classification

The performance of the validated (depth = 3) deep learning framework on pair-wise (binary) classification tasks was compared to identify how well the pMCI and AD populations separated from the CN and sMCI populations. These binary classification tasks were conducted using repeated (n = 10), stratified 5-fold cross-validation procedures and model training was conducted with an early stopping with a patience level of 20 epochs (20% of the set maximum number of epochs) to prevent overtraining the model. The results in Figure 4 reflect a clear trend with the average (cross-validated over 50 folds) classification metrics for the classification of CN or sMCI classes from pMCI or AD classes distinctly higher than the average metrics for the CN vs. sMCI and pMCI vs. AD classification tasks. Specifically, for the first four classification tasks (CN vs. AD, CN vs. pMCI, sMCI vs. AD, and sMCI vs. pMCI), we report a cross-validated median accuracy of 91%, 86%, 86% and 77% respectively. The reported sensitivity for these tasks is 85%, 79%, 81% and 71% respectively, whereas the reported specificity is 95%, 92%, 90% and 82% respectively. The appropriate separability trend across the different classes and genuinely high classification metrics as compared to previous findings in literature (reviewed recently in (Moradi et al. 2015 and Vieira et al., 2017) in such a large heterogenous sample clearly highlight the usefulness of the used deep learning model.

For further introspection into the diagnostic ability of the binary classifiers, we estimated the classification-task-specific receiver operating characteristic (ROC) curves. A comparison of the area under the ROC curve (AUC) metric confirmed a similar trend as suggested in the previous analysis (Figure 4) as illustrated in Figure 5. We report a cross-validated AUC of 0.93 for CN vs. AD, 0.87 for CN vs. pMCI, 0.89 for sMCI vs. AD and 0.81 for the sMCI vs. pMCI classification tasks. These initial results indicate high suitability of the evaluated framework for our desired objective; further possible improvement in prediction of progression to AD was explored with the mixed-class

prognostic classification analysis as discussed in the next section. This is followed by a thorough, in-depth comparison of our prediction performance with previous literature.

Mixed-class Prognostic Classification

The sMCI vs. pMCI classification task could be considered as the most clinically relevant task amongst the several binary classification tasks since identifying MCI subjects who are highly likely to progress to AD is very crucial; hence, in this specific analysis we focus on exploring ways to improve separability between these two classes. A recent study (Cheng et al. 2015) explored advantages of domain transfer learning to enhance MCI conversion prediction rates which is similar to what we pursue in the section. In general, training the learner with more data is highly likely to improve its classification/prediction performance on unseen data since the learning model assimilates the additional variability provided by the previously unseen datasets and adjusts its weights accordingly for more generalized training (i.e. decrease in generalization error). In a scenario where availability of MCI data is severely limited, we hypothesized that training the learner with all data from the CN and AD classes (or domains) together with some part of the two MCI classes (or domains), and then testing with the remaining part of the MCI classes (or domains) could enhance classification performance. For this analysis, we conducted the above discussed modified form of repeated (n=10) stratified 5-fold cross-validation and report significantly improved cross-validated median accuracy of 83%, sensitivity of 78%, and specificity of 87% (Figure 6A), and a cross-validated mean AUC of 0.88 (Figure 6B). The results clearly reflect substantial improvement (6% in accuracy, 7% in sensitivity, 5% in specificity and 7% in AUC) with the addition of domain transfer learning in the training phase. Finally, the performance of this modified inter-MCI case was confirmed as a significant improvement over a standard machine learning approach such as the classical support vector machine (SVM) classifier (p = 2.5762 x 10^{-8}) applied on the same training/testing cross-validation folds . In this specific analysis, for estimating the performance of the SVM classifier, the classical univariate feature selection procedure using F-

test (ANOVA) was implemented for dimension reduction following which the optimal value of the penalty (cost) parameter in the linear SVM was estimated. The boxplots for the accuracies for the different cross-validation folds using the Resnet and SVM models are shown in Figure 6C.

Comparison with previous literature

In this section, we compare the prediction performance of AD progression in our study (modified inter-MCI task) to previous deep learning work in recent literature (Table 2). In order to identify previous studies that used deep learning on neuroimaging data to study psychiatric or neurological disorders, we conducted a search on PubMed (May 25, 2018) using search terms very similar to a recent review (Vieira, Pinaya, and Mechelli 2017). Specifically, the following search terms were used: ("deep learning" OR "deep architecture" OR "artificial neural network" OR "autoencoder" OR "convolutional neural network" OR "deep belief network") AND (neurology OR neurological OR psychiatry OR psychiatric OR diagnosis OR prediction OR prognosis OR outcome) AND (neuroimaging OR MRI OR "magnetic resonance imaging" OR "fMRI" OR "functional magnetic resonance imaging" OR PET OR "positron emission tomography"). Following this, we manually screened these articles to identify the relevant subset of studies that applied deep learning to study MCI to AD progression. A comparison of prediction using MRI data only confirms the superior performance of our method as compared to other undertaken approaches. Using just MRI data, the prediction accuracy obtained in our study (82.7%) is more than 7% greater than the second best performer (using MRI data only) that used a multiscale deep NN in a very recent study (Lu et al. 2018). Considering use of multiple modalities, only H. Il Suk et al., 2015 (83.3% using MRI, PET and CSF modalities) and Lu et al., 2018 (82.93% using MRI and PET) report slightly higher performance as compared to our study. Interestingly, despite using multiple modalities, the methods used in these two particular studies report only marginal improvements (0.6% and 0.2% respectively) over our unimodal analysis. Working with multiple modalities generally enhances the prediction performance (variably from 3% to greater than 20% in studies included in Table 2), so it would be

reasonable to expect further improvement in prediction performance through our method if complimentary information from an additional modality is leveraged.

Furthermore, Moradi et al., 2015 (see Table 7 in their manuscript) and Korolev et al., 2016 (see Table3 in their manuscript) did extensive comparisons of other (non-deep-learning) studies and showed their respective approaches to result in better precision than other approaches in previous literature. Moradi et al. 2015 studied progression with ADNI data (large sample of 825 subjects) using a regularized logistic regression approach to report classification accuracy of 74% using MRI biomarker only and 82% using their aggregate biomarker that used the patient age and clinical scores as features in addition to the MRI biomarker. Korolev et al. 2016 worked with only ADNI-1 MCI subjects (n = 259) to predict progression to AD from MCI using a probabilistic, kernel-based pattern classification approach to report a prediction accuracy of 79.9% using MRI and clinical (cognitive and functional) scores. Our method predicted more accurately (82.7%) using a large sample of 828 subjects (MRI data alone) than these two multimodal, non-deep-learning studies and all studies reviewed in these two studies.

Multi-class (4-way) Diagnostic/Prognostic Classification

For the multi-class (4-way) case, the learning framework scored a cross-validated median accuracy of 54% that is higher than recent studies evaluating such a 4-way classification (as reviewed in Table 2 in Vieira, Pinaya, and Mechelli 2017). Although this accuracy level is substantially higher than chance (25%), the appropriateness of the data trends learnt in this much harder classification problem was further confirmed by in-depth ROC and feature projection analyses as discussed next. As an extension of binary ROC analysis, for each class, we estimated a single ROC curve by comparing it to all other classes (i.e. one vs all comparison). ROC curves for the multi-class case can also be estimated by micro-averaging which measures true and false positive rates by considering each element of each class as a binary prediction, or by macro-averaging which essentially averages

over the several class-specific classification metrics. In this analysis, the AD and CN classes reported a higher AUC followed by the micro-averaged and macro-average cases, whereas the pMCI and sMCI classes showed lower AUC (Figure 7A).

In the multi-class feature projection analyses (Figure 7B and 7C), the 512-dimensional features at the output of the first fully-connected layer in the employed framework were projected onto a two-dimensional space using the tSNE algorithm (der Maaten et al. 2008). The tSNE algorithm embeds similar observations as nearby points and non-similar observations as distant points with high probability; so more similar classes could be expected to cluster in vicinity of each other in the projection space. This projection analysis was performed to confirm the learning directionality of validated models in our multi-class classification case expecting majority observations for more similar classes being projected/clustered together. Figure 7B demonstrates projections from a sample surrogate model (i.e. model validated for a sample cross-validation fold). Although the classes are not separable in the projection space, yet a clear pattern can be traced easily across the projection spectrum. More specifically, we can observe classes ordered in increasing severity of disease from left to right (i.e. CN, sMCI, pMCI and AD in this specific order) although some outlier observations do exist. The disease severity or the class pattern is further confirmed by coloring the same two-dimensional projections (as in Figure 7B) with the six clinical (cognitive and functional) scores (Figure 7C). The MMSE and RAVLT clinical scores reveal a clear increase across the spectrum (left through right), whereas the FAQ, CDRSB, ADAS11 and ADAS13 clinical scores (by nature of score characterization) reveal a clear decrease across the same spectrum.

Model Validation for the Mixed Class (modified inter-MCI) Classification Task

The mixed class (i.e. modified sMCI vs. pMCI) binary classification task was focussed on to identify the most discriminative brain regions in early AD. To ensure proper learning in the training phase, additional tests were performed to filter weights and output activations at the first convolutional

layer. As in the case of any CNN architecture, similar (i.e. highly correlated) filters almost certainly indicate incorrect convergence of the loss function and thus an inappropriate local minimum. To control for that, we performed a similarity analysis on the weights of the 64 (size 3 x 3 x 3) filters learnt at the first convolution layer as a sanity check. More specifically, the filter weights for each of the channels in the first convolutional layer were compared pair-wise in each of the 50 surrogate models as well as the final predictive model. Boxplots in columns 1 to 50 in Figure 8A show the pair-wise correlations of filter weights for each of the 50 surroagte models, whereas those for the final predictive model are shown in the boxplot in column 51. Since low pair-wise correlations in the learnt 64 filters in each of these 51 models were observed, it can be concluded that the trained model indeed learnt the local minima.

While inference on discriminative brain regions was done from the features of the final predictive models that were trained on all data, ideally, projections of data samples into the representation space estimated by the surrogate models must be similar to confirm similar learning trends and robustness of the extracted features. For this comparison, the final predictive model and the 50 surrogate models for this classification task were compared in the mean activations across the 64 channels at the output of the first convolutional layer (batch-normalized and post non-linear activation). A pairwise similarity (i.e. correlation) comparison of the 3D activation maps across these 51 models is shown on the left in Figure 8B, whereas the comparison of the similarity in activation maps for the 50 surrogate models to the activation maps from the final predictive model is shown on the right in the same panel. In both cases, a very high median correlation value was observed thus confirming similar discriminative features extracted across different surrogate models and the final predictive model.

Localizing Abnormalities: Discriminative Brain Regions

Peak activations of the identified brain regions which are most discriminative of progression of MCI to AD were localized by projecting the sub-sampled mean activations at the output of the first convolutional layer (batch-normalized and post non-linear activation) back to the brain space. Notably, the activations at the output of the first convolutional layer could be expected to be fine-tuned to small degrees in the deeper layers. However, we restricted our analysis to the first layer since there is a substantial loss of spatial resolution as we go deeper into the networks, and estimating and back-propagating the relative importance of the features at each layer needs dedicated developmental efforts that are outside the scope of the current work. Hence, we restrict our abnormality localization analysis to features at the output of the first convolutional layer. The brain regions at the peak activations were identified in correspondance to the AAL brain atlas. Figure 8C illustrates the spatial maps at the most activated saggital, coronal and axial slices as tracked from the final predictive model for the modified sMCI vs. pMCI classification task. In this figure, hotter (towards the red spectrum) colors imply higher mean activations, while cooler (towards the blue spectrum) colors imply lower mean activations. Specifically, peak activations were observed in the sub-cortical regions including hippocampus, amygdala, caudate nucleus, putamen and thalamus, cortical regions including anterior cingulate gyrus and middle cingulate gyrus, temporal regions including fusiform gyrus, temporal pole, and temporal inferior gyrus, frontal superior/middle gyrus, cerebellum, parietal regions including angular gyrus, precuneus and supramarginal gyrus, and occipital regions including calcarine, cuneus and lingual gyrus. In the discussion section, we will discuss these discriminative brain regions in correspondance to several previously reported findings in the AD/MCI literature.

DISCUSSION

In this work, we extensively test the ability of the ResNets to learn abstract neuroanatomical alterations in structural MRI data. The tested deep learning architecture provided close to the state of the art prognostic classification performance following which we focused on the inter-MCI classification task to predict the MCI sub-group who would progress to AD within three years. The principle progression analysis of our work is the mixed-class (i.e. modified) inter-MCI classification task where we used principles of domain transfer learning (additionally training with data from other domains). This particular analysis bears high clinical relevance. Importantly, on the MRI data alone we achieved classification accuracy of 82.7% which is a significant improvement over state of the art with either MRI based (75.44% as reported in Lu et al., 2018) and very close to state of the art performance with multimodal results (83.3% as reported in H. Il Suk et al., 2015). The accuracy in this modified inter-MCI class classification task is significantly higher than that in the standard inter-MCI case which suggests the performance improvement was also enabled by additional training information acquired from other (AD and CN) domains. Furthermore, the learning directionality and trends were verified in the multiclass case by projecting the features at output of the first fully-connected layer onto a two-dimensional surface. The projection/clustering class sequence in Figure 7B and 7C support the appropriateness of the extracted features and their association with the clinical scores, thus confirming the high learning capacity and potential of this deep architecture. The learning directionality and trends were further validated by additionally testing for similarity in the filter weights and activations at the first convolutional layer (Figure 8A and 8B) for the different surrogate models and the final predictive model. Notably, the reported performance metrics were obtained from a large dataset (n = 828), a rigorous cross-validation procedure featuring 10 repeats and a sufficiently large (20%) test size in each of the folds (i.e. a total number of 50 large-sized stratified folds). These results evince that the ResNets can be considered well-suited to neuroimaging data and future studies to uncover further potential of such

or similar architectures must be undertaken. Next, we discuss the discriminative brain regions suggested by the ResNet in context to previous findings in literature.

Discriminative brain regions

AD is characterized by serious trouble in performing familiar tasks, solving problems, planning, reasoning, judgement and thinking, and generally features increased confusion and discomfort in speech, vision, reading, focusing, and spatial or temporal perception. Struggling with these symptoms, the person undergoes mood and personality changes and increasingly loses interest in favorite activities and social life. A sizable amount of previous work has related the above mentioned cognitive, behavioral and emotional phenomenon to specific structural changes in the brain, which we discuss next in particular context to the discriminative brain regions identified by the ResNet framework.

The hippocampus and amygdala subcortical regions in the medial temporal lobe have been consistently reported as most prominent discriminative regions in early AD. Hippocampus is strongly related to memory formation and recall, and recent evidence suggests more pronounced hippocampal atrophy in the progressive MCI class (Braak and Braak 1991b; Burton et al. 2009; Costafreda et al. 2011; Devanand et al. 2007; Kantarci et al. 2009; Risacher et al. 2009; Visser et al. 2002; Walhovd et al. 2010). Similarly, structural changes in amygdala, a brain region mainly responsible for emotional experiences and expressions, have been related to personality changes, for example, increased irritability and anxiety, in AD (Poulin et al. 2011; Unger et al. 1991; Whitwell et al. 2008). Other highly activated subcortical regions included thalamus and the dorsal striatum (putamen and caudate). While the main function of thalamus is to relay motor and sensory signals to the cerebral cortex, and regulate consciousness and sleep, the dorsal striatum is believed to contribute directly to decision-making subjective to desired goals. Observed aberrations in the dorsal striatum (putamen and caudate) and thalamus regions are typical of subjects with AD

(Aggleton et al. 2016; Braak and Braak 1991a; Cho et al. 2014; Jiji et al. 2013; De Jong et al. 2008) with impairments in the thalamus in AD associated to deteriorating consciousness, bodily movement and coordination, attentional, and motivation levels and impairments in the dorsal striatum associated to very slow or absent decision-making abilities.

Apart from the above widely studied and highly discriminative medial temporal lobe, we also report peak activations in the fusiform gyrus, inferior temporal gyrus and temporal pole regions on the temporal lobe. These regions have been known to be associated with pattern (e.g. face, body, object, word, color, etc.) recognition and reported to be affected by AD in a few previous studies (Chan et al. 2001; Galton et al. 2011). In the frontal lobe, peak activations were observed in the superior and middle frontal gyrus. These regions are also associated with decision making and problem solving, reportedly highly damaged in AD (Johnson et al. 2005; Sluimer et al. 2009; Whitwell et al. 2008) and are believed to lead to higher lethargy levels, bizarre/inappropriate behavior and situations of being stuck in a certain condition (repeating same things over and over again).

Besides the above discussed frontotemporal networks, AD is characterized by decline in important parietal networks such as precuneus (Apostolova and Thompson 2008; Bailly et al. 2015; Fennema-Notestine et al. 2009; Scahill et al. 2002; Walhovd et al. 2010; Whitwell et al. 2008), and the angular and supramarginal gyrus regions (Walhovd et al. 2010; Yun, Kwak, and Lee 2015). Cerebellum, a critical brain region in several motor, cognitive and behavioral functions, is also more recently being increasingly suggested as a direct contributor to cognitive and neuropsychiatric deficits in AD (Guo et al. 2016; Jacobs et al. 2017; Schmahmann 2016) with deteriorating cerebellum health resulting in several symptoms such as lack of balance and coordination, tremors, slurred speech and abnormal eye movements in the elderly.

Cortical regions including anterior cingulate gyrus and middle cingulate gyrus that are primarily responsible for higher cognitive (i.e. decision-making) and emotional (e.g. social interactions,

empathy, etc.) processes have also been suggested to be highly vulnerable in AD (Bailly et al. 2015; Fennema-Notestine et al. 2009; Huang et al. 2002; Jones et al. 2006; Killiany et al. 2000; Scahill et al. 2002). Finally, damages to the occipital lobe are associated with increased misinterpretations of the surrounding environment (e.g. hallucinations, illusions, misidentification, misperceptions, etc.) and occipital regions comprising the calcarine, cuneus and lingual gyrus regions have indeed been reported to be compromised in progression to AD.

The above discussed findings add further evidence that the localized abnormal patterns in the brain structure could play a significant role in prediction of early AD biomarkers and are of potential clinical application. In fact, a few of the discriminative regions that we report are rarely used as prognostic biomarkers to study conversion of MCI to AD; our work and the citied literature in this particular discussion provide a compelling evidence of including these additional biomarkers to allow for a complete characterization of the structural changes in AD progression.

Limitations and future scope

Here we note some basic limitations of our work that could be addressed in the future depending on algorithmic computational tractability, and availability of data resources and data processing algorithms. As with other neuroimaging studies, the foremost limitation is a limited training data size. In generic image processing applications, this limitation is often addressed with data augmentation procedures by using simple rotation, translation, scaling and other data transformations (also see Castro et al., 2015 and Ulloa et al., 2015 for more elaborate data augmentation examples with structural MRI). We expect even further increases in performance with employing such techniques in the future work. This broadens perspectives for our models that are already performing at or above the state of the art. Interestingly, a recent study (Casanova, Hsu, and Espeland 2012) demonstrated an increase in classification performance with increase in sample size using ADNI structural MRI data. Similarly, in our work as well, we saw a substantial

increase in performance with more training data being fed to the ResNet framework in the modified inter-MCI class classification task as compared to the standard inter-MCI class classification task. This makes a strong case to test use of multiple datasets to extract features in a pooled or separate fashion, and then use the pooled or separate information to train the machine learning framework. With increasing data availability and standardization in data preprocessing and pooling procedures, further substantial improvement in diagnostic and/or prognostic classification performance could be expected in future multi-study deep learning research efforts.

Due to the computationally expensive nature of training deep CNNs, few limitations in regards to computational tractability within realistic study time remain. This tends to restrict extensive fine-tuning of each involved hyperparameter through random or grid search analysis on multiple hyperparameters and additionally backing up statistical trends using methods such as Monte-Carlo. As such, the most important hyperparameters must be prioritized and optimized to estimate generic performance trends of the algorithm within the realistic study period. For this specific work, we optimized the initial learning rate and L2 weight decay parameter on a sample cross-validation fold using extensive grid analysis, and retained the values for other dataset partitions. Although the same hyperparameters would likely achieve close to actual performance on other data folds, yet this fine-tuning could have a small effect on the performance of the respective surrogate models (e.g. reported performance metrics could be slightly lower than the original) but also on that of the final predictive model. It must be noted that this limitation is for performance quantification only; it is least likely to affect the qualitative analysis (e.g. localizing discriminative brain regions) by a significant margin.

Choosing stopping criterion for learning a classifier typically involves a tradeoff between generalization error and learning time. While this study approximated the stopping criterion with information across all cross-validation folds, further detailed introspection using relatively unestablished but promising variants of early stopping criterion could be explored in future

investigations (Prechelt 1998). Similarly, effect of algorithmic variations in bottleneck residual block structures (size and depth), training time, and loss optimization procedures could be understood in future studies to further enhance the prediction performance. Finally, while we focus on activations at the output of the first convolutional layer, we could expect sequential fine-tuning of such activations as we move down in the architecture. This however happens at the cost of substantial loss of spatial resolution. Independent frameworks are now being proposed to study multi-layer relevance propagation (Bach et al. 2015; Lapuschkin et al. 2016) in 2D image processing applications using deep CNN or ResNet frameworks. In fact, multi-layer relevance propagation is currently an independent, rich research topic as it is equally important to understand/interpret the predictions as predicting itself, and even more so in medical applications. However, interpreting these deep non-linear models and optimally representing extracted features is not trivial. In absence of a framework to allow for fine-tuned optimal representations of the features/activations in the used architecture, we restricted our analysis to the activations at the first convolutional layer as these representations could be considered as a good representation of the discriminative features that are further fine-tuned to small degrees at each layer. Nonetheless, it would be interesting to explore any significant activation changes with more deep relevance propagation frameworks once they become available.

Several other approaches for enhancing predictive performance of AD progression could be explored in future work. Diagnosis for the subjects is currently established through clinical scores but diagnosis-specific neuroanatomical or neurofunctional abnormalities might actually not show in each subject in each class due to the heterogeneous nature of age-related dementia. In such a scenario, it could be interesting to constrain this heterogeneity by training the machine learning model on the most homogeneous samples (i.e. samples most representative of the given class) and then evaluate change in the performance of the diagnostic/prognostic classification or change in the feature space of interest. Another approach could be to fuse the low-dimensional clinical scores

used to make the clinical diagnosis with the MRI features space to further enrich feature learning process. This approach has reportedly resulted in enhanced performance in few studies as also suggested in Table 2. Other widely used low dimensional features chosen by experts (e.g. volumetric MRI features or similar features from other modalities) could also further enhance diagnostic/prognostic performance.

Recent literature reflects ample evidence of advantages of multimodal studies in understanding brain structure and/or function and decode brain complexities (Abrol, Rashid, et al. 2017; Calhoun and Adali 2009; Calhoun and Sui 2016). Indeed, few previous multimodal studies have reported prediction performance improvements due to training the same machine learning framework with multiple modalities as compared to a single studied modality for studying AD/MCI (Lorenzi et al. 2016; Toledo et al. 2013; Zhang et al. 2011) as is also evident from Table 2. Due to this evidence from other explored machine learning neuroimaging studies, a performance improvement is highly likely if features for multiple modalities are extracted through the ResNet framework and further fused using a data fusion algorithm to generate a collective feature space for predicting chances of progression to AD.

Interestingly, fusion of features from multiple structural (MRI, PET and CSF) modalities (structure-structure fusion) has been much more frequently explored than fusion of feature space from one or more of these structural modalities to feature space from a functional neuroimaging modality (structure-function fusion) such as fMRI. One of the reasons for the relatively less explored structure-function fusion in AD/MCI literature could be the significantly smaller number of fMRI datasets as compared to data from the structural modalities. Nonetheless, structure-function fusion could be highly interesting and several robust fMRI features such as amplitude of low frequency fluctuation (ALFF) maps, or static/time-varying functional connectivity (FC) maps exist. Of specific interest in fMRI is the time-varying FC feature space that have recently been shown to be replicable (Abrol et al. 2016), statistically significant and robust against variation in data grouping,

decomposition and analysis methods (Abrol, Damaraju, et al. 2017), and also more discriminative of diseased brain conditions (Rashid et al. 2016) than static FC. As such, future works featuring such promising deep learning models could seek performance gains not only from structure-structure fusion coupled with information in cognitive/functional scores but with structure-function fusion as well.

Conclusion

This work shows that the ResNet architecture showed performance beyond the state of the art in predicting progression to AD using MRI data alone, and within 1% of the state of art performance considering multimodal studies as well. This clearly reflects the high potential of this deep architecture for studying progression to AD and neuroimaging data in a broader sense. The prognostic classification performance was exceptional despite several limitations as outlined in the discussion section and addressing these limitations in future work could highly likely result in further improvement in performance of this relatively newer machine learning framework. The most discriminative brain regions as highlighted by the ResNet framework confer with previous findings in AD/MCI literature to a high degree, and regions for which there is insufficient evidence must be investigated further to enhance the set of potential AD biomarkers. The ResNet architecture could be explored in future for learning from multiple modalities for investigating any possible improvements in diagnostic and prognostic classification and identification of more specific multi-modal biomarkers for AD or other brain conditions. We conclude that our results further strengthen the expectations and high likelihood of discovery of modifiable risk factors for understanding biomarkers of progression to AD early, especially through use of advanced neuroimaging data processing methods such as the one explored in this work.

ACKNOWLEDGEMENT

This work was supported by NIH grant numbers 2R01EB005846, P20GM103472, R01REB020407 and R01EB006841 as well as NSF grant 1539067 to Dr. Vince D. Calhoun, National Natural Science Foundation of China grant 61703253, and Natural Science Foundation of Shanxi Province in China grant 2016021077 to Dr. Yuhui Du, and NSF grant IIS-1318759 to Dr. Sergey Plis.

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.

REFERENCES

Abrol, A., E. Damaraju, et al. 2017. "Replicability of Time-Varying Connectivity Patterns in Large Resting State FMRI Samples." *NeuroImage* 163.

Abrol, A., B. Rashid, et al. 2017. "Schizophrenia Shows Disrupted Links between Brain Volume and Dynamic Functional Connectivity." *Frontiers in Neuroscience* 11(NOV).

Abrol, A., C. Chaze, E. Damaraju, and V.D. Calhoun. 2016. "The Chronnectome: Evaluating Replicability of Dynamic Connectivity Patterns in 7500 Resting FMRI Datasets." In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*,.

"ADNI MRI Protocols." 2017. http://adni.loni.usc.edu/methods/documents/mri-protocols (July 20, 2005).

Aggleton, John P., Agathe Pralus, Andrew J.D. Nelson, and Michael Hornberger. 2016. "Thalamic Pathology and Memory Loss in Early Alzheimer's Disease: Moving the Focus from the Medial Temporal Lobe to Papez Circuit." *Brain* 139(7): 1877–90.

Apostolova, Liana G., and Paul M. Thompson. 2008. "Mapping Progressive Brain Structural Changes in Early Alzheimer's Disease and Mild Cognitive Impairment." *Neuropsychologia* 46(6): 1597–1612.

Arbabshirani, Mohammad R., Sergey Plis, Jing Sui, and Vince D. Calhoun. 2017. "Single Subject Prediction of Brain Disorders in Neuroimaging: Promises and Pitfalls." *NeuroImage* 145: 137–65.

Bach, Sebastian et al. 2015. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PLoS ONE* 10(7).

Bailly, Matthieu et al. 2015. "Precuneus and Cingulate Cortex Atrophy and Hypometabolism in Patients with Alzheimer's Disease and Mild Cognitive Impairment: MRI and 18 F-FDG PET Quantitative Analysis Using FreeSurfer." *BioMed Research International* 2015(Mci): 1–8.

Braak, H., and E. Braak. 1991a. "Alzheimer's Disease Affects Limbic Nuclei of the Thalamus." *Acta Neuropathologica* 81: 261–68.

———. 1991b. "Neuropathological Stageing of Alzheimer-Related Changes." *Acta Neuropathologica* 82(4): 239–59.

Bradford, Andrea et al. 2009. "Missed and Delayed Diagnosis of Dementia in Primary Care." *Alzheimer Disease & Associated Disorders* 23(4): 306–14.

Burton, E J et al. 2009. "Medial Temporal Lobe Atrophy on MRI Differentiates Alzheimer's Disease from Dementia with Lewy Bodies and Vascular Cognitive Impairment: A Prospective Study with Pathological Verification of Diagnosis." *Brain* 132(1): 195–203.

Calhoun, Vince D., and Tülay Adali. 2009. "Feature-Based Fusion of Medical Imaging Data." *IEEE Transactions on Information Technology in Biomedicine* 13(5): 711–20.

Calhoun, Vince D., and Jing Sui. 2016. "Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness." *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 1(3): 230–44.

Casanova, Ramon, Fang Chi Hsu, and Mark A. Espeland. 2012. "Classification of Structural MRI Images in Alzheimer's Disease from the Perspective of Ill-Posed Problems." *PLoS ONE* 7(10).

Casey, David a, Demetra Antimisiaris, and James O'Brien. 2010. "Drugs for Alzheimer's Disease: Are They Effective?" *P & T : a peer-reviewed journal for formulary management* 35(4): 208–11.

Castro, E. et al. 2015. "Generation of Synthetic Structural Magnetic Resonance Images for Deep Learning Pre-Training." In *Proceedings - International Symposium on Biomedical Imaging*, , 1057–60.

Chan, Dennis et al. 2001. "Patterns of Temporal Lobe Atrophy in Semantic Dementia and Alzheimer's Disease." *Annals of Neurology* 49(4): 433–42.

Chen, Yani, Bibo Shi, Charles D. Smith, and Jundong Liu. 2015. "Nonlinear Feature Transformation and Deep Fusion for Alzheimer's Disease Staging Analysis." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, , 304–12.

Cheng, Bo, Mingxia Liu, Daoqiang Zhang, and Brent C Munsell. 2015. "Domain Transfer Learning for MCI Coversion Prediction." *Ieee Transaction on Biomedical Engineering* 62(7): 1805–17.

Cho, Hanna et al. 2014. "Shape Changes of the Basal Ganglia and Thalamus in Alzheimer's Disease: A Three-Year Longitudinal Study." *Journal of Alzheimer's Disease* 40(2): 285–95.

Connolly, Amanda et al. 2011. "Underdiagnosis of Dementia in Primary Care: Variations in the Observed Prevalence and Comparisons to the Expected Prevalence." *Aging and Mental Health* 15(8): 978–84.

Costafreda, Sergi G. et al. 2011. "Automated Hippocampal Shape Analysis Predicts the Onset of Dementia in Mild Cognitive Impairment." *NeuroImage* 56(1): 212–19.

Devanand, D. P. et al. 2007. "Hippocampal and Entorhinal Atrophy in Mild Cognitive Impairment: Prediction of Alzheimer Disease." *Neurology* 68(11): 828–36.

Falahati, Farshad, Eric Westman, and Andrew Simmons. 2014. "Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging." *Journal of Alzheimer's Disease* 41(3): 685–708.

Fennema-Notestine, Christine et al. 2009. "Structural MRI Biomarkers for Preclinical and Mild Alzheimer's Disease." *Human Brain Mapping* 30(10): 3238–53.

"GAAIN Data." 2017. https://www.gaaindata.org/partners/online.html (August 20, 2005).

Galton, C J et al. 2011. "Differing Patterns of Temporal Atrophy in Alzheimer ' s Disease." *New York* 10(4): 220–25.

Guo, Christine C. et al. 2016. "Network-Selective Vulnerability of the Human Cerebellum to Alzheimer's Disease and Frontotemporal Dementia." *Brain* 139(5): 1527–38.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. "Deep Residual Learning for Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, , 770–78.

———. 2016b. "Identity Mappings in Deep Residual Networks." In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), , 630–45.

Huang, Chaorui et al. 2002. "Cingulate Cortex Hypoperfusion Predicts Alzheimer's Disease in Mild Cognitive Impairment." *BMC Neurology* 2(1): 9.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *CoRR* abs/1502.0.

Jacobs, Heidi I. L. et al. 2017. "The Cerebellum in Alzheimer's Disease: Evaluating Its Role in Cognitive Decline." *Brain*.

Jiji, Sudevan et al. 2013. "Segmentation and Volumetric Analysis of the Caudate Nucleus in Alzheimer's Disease." *European Journal of Radiology* 82(9): 1525–30.

Johnson, Nathan A. et al. 2005. "Pattern of Cerebral Hypoperfusion in Alzheimer Disease and Mild Cognitive Impairment Measured with Arterial Spin-Labeling MR Imaging: Initial Experience." *Radiology* 234(3): 851–59. http://pubs.rsna.org/doi/10.1148/radiol.2343040197.

Jones, Bethany F. et al. 2006. "Differential Regional Atrophy of the Cingulate Gyrus in Alzheimer Disease: A Volumetric MRI Study." *Cerebral Cortex* 16(12): 1701–8.

De Jong, L. W. et al. 2008. "Strongly Reduced Volumes of Putamen and Thalamus in Alzheimer's Disease: An MRI Study." *Brain* 131(12): 3277–85.

Kantarci, K. et al. 2009. "Risk of Dementia in MCI: Combined Effect of Cerebrovascular Disease, Volumetric MRI, and 1H MRS." *Neurology* 72(17): 1519–25.

Killiany, R J et al. 2000. "Use of Structural Magnetic Resonance Imaging to Predict Who Will Get Alzheimer's Disease." *Annals of neurology* 47(4): 430–39.

Kingma, Diederik P., and Jimmy Lei Ba. 2015. "Adam: A Method for Stochastic Optimization." *International Conference on Learning Representations 2015*: 1–15.

Korolev, Igor O, Laura L Symonds, Andrea C Bozoki, and Alzheimer's Disease Neuroimaging Alzheimer's Disease Neuroimaging Initiative. 2016. "Predicting Progression from Mild Cognitive Impairment to Alzheimer's Dementia Using Clinical, MRI, and Plasma Biomarkers via Probabilistic Pattern Classification." *PloS one* 11(2): e0138866.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances In Neural Information Processing Systems*: 1–9.

Lapuschkin, Sebastian et al. 2016. "The Layer-Wise Relevance Propagation Toolbox for Artificial Neural Networks." *Journal of Machine Learning Research* 17(114): 1–5. http://jmlr.org/papers/v17/15-618.html.

Li, Feng et al. 2015. "A Robust Deep Model for Improved Classification of AD/MCI Patients." *IEEE Journal of Biomedical and Health Informatics* 19(5): 1610–16.

Litjens, Geert et al. 2017. "A Survey on Deep Learning in Medical Image Analysis." *Medical Image Analysis* 42: 60–88.

Liu, Manhua, Daoqiang Zhang, and Dinggang Shen. 2014. "Hierarchical Fusion of Features and Classifier Decisions for Alzheimer's Disease Diagnosis." *Human Brain Mapping* 35(4): 1305–19.

Liu, Siqi et al. 2015. "Multi-Phase Feature Representation Learning for Neurodegenerative Disease Diagnosis." In *1st Australasian Conference on Artificial Life and Computational Intelligence, ACALCI 2015, February 5, 2015 - February 7*, , 350–59.

Lorenzi, Marco et al. 2016. "Multimodal Image Analysis in Alzheimer's Disease via Statistical Modelling of Non-Local Intensity Correlations." *Scientific Reports* 6.

Lu, Donghuan et al. 2018. "Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease Using Structural MR and FDG-PET Images." *Scientific Reports* 8(1): 5697. https://doi.org/10.1038/s41598-018-22871-z.

der Maaten, Laurens et al. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9(2579–2605): 85.

Markesbery, William R. 2010. "Neuropathologic Alterations in Mild Cognitive Impairment: A Review." *Journal of Alzheimer's Disease* 19(1): 221–28.

Moradi, Elaheh et al. 2015. "Machine Learning Framework for Early MRI-Based Alzheimer's Conversion Prediction in MCI Subjects." *NeuroImage* 104: 398–412.

Plis, Sergey M. et al. 2014. "Deep Learning for Neuroimaging: A Validation Study." *Frontiers in Neuroscience* (8 JUL).

Poulin, Stéphane P. et al. 2011. "Amygdala Atrophy Is Prominent in Early Alzheimer's Disease and Relates to Symptom Severity." *Psychiatry Research - Neuroimaging* 194(1): 7–13.

Prechelt, L. 1998. "Early Stopping--But When?" In *Neural Networks: Tricks of the Trade*, , 55–69.

"Pytorch Resnet Architecture." 2017. https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py (March 7, 2018).

Rashid, Barnaly et al. 2016. "Classification of Schizophrenia and Bipolar Patients Using Static and Dynamic Resting-State FMRI Brain Connectivity." *NeuroImage* 134: 645–57.

Rathore, Saima et al. 2017. "A Review on Neuroimaging-Based Classification Studies and Associated Feature Extraction Methods for Alzheimer's Disease and Its Prodromal Stages." *NeuroImage* 155: 530–48.

Risacher, Shannon et al. 2009. "Baseline MRI Predictors of Conversion from MCI to Probable AD in the ADNI Cohort." *Current Alzheimer Research* 6(4): 347–61.

Scahill, R. I. et al. 2002. "Mapping the Evolution of Regional Atrophy in Alzheimer's Disease: Unbiased Analysis of Fluid-Registered Serial MRI." *Proceedings of the National Academy of Sciences* 99(7): 4703–7.

Schmahmann, Jeremy D. 2016. "Cerebellum in Alzheimer's Disease and Frontotemporal Dementia: Not a Silent Bystander." *Brain* 139(5): 1314–18.

Shen, Dinggang, Guorong Wu, and Heung-Il Suk. 2017. "Deep Learning in Medical Image Analysis." *Annual Review of Biomedical Engineering* 19(1): 221–48.

Shi, Jun et al. 2018. "Multimodal Neuroimaging Feature Learning with Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease." *IEEE Journal of Biomedical and Health Informatics* 22(1): 173–83.

Simonyan, Karen, and Andrew Zisserman. 2015. "VGG: Very Deep Convolutional Networks for Large-Scale Image Recognition." *International Conference on Learning Representations (ICRL)*: 1–14.

Sluimer, Jasper D. et al. 2009. "Accelerating Regional Atrophy Rates in the Progression from Normal Aging to Alzheimer's Disease." *European Radiology* 19(12): 2826–33.

Srivastava, Rupesh Kumar, Klaus Greff, and Jürgen Schmidhuber. 2015. "Training Very Deep Networks." *NIPS*: 1–9.

Suk, Heung-Il, Seong-Whan Lee, and Dinggang Shen. 2015. "Deep Sparse Multi-Task Learning for Feature Selection in Alzheimer's Disease Diagnosis." *Brain structure & function*.

Suk, Heung-Il, Seong-Whan Lee, Dinggang Shen, and for the Alzheimer's Disease Neuroimaging Initiative. 2017. "Deep Ensemble Learning of Sparse Regression Models for Brain Disease Diagnosis." *Medical Image Analysis* (In Press, Accepted Manuscript).

Suk, Heung Il, Seong Whan Lee, and Dinggang Shen. 2014. "Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis." *NeuroImage* 101: 569–82.

———. 2015. "Latent Feature Representation with Stacked Auto-Encoder for AD/MCI Diagnosis." *Brain Structure and Function* 220(2): 841–59.

Suk, Heung Il, and Dinggang Shen. 2013a. "Deep Learning-Based Feature Representation for AD/MCI Classification." In Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), , 583–90.

———. 2013b. "Deep Learning Based Feature Representation for AD/MCI Classification." In *Medical Image Computing and Computer-Assisted Intervention (MICCAI), 16th International Conference*, , 583–90.

Szegedy, Christian et al. 2015. "Going Deeper with Convolutions(Inception, GoogLeNet)." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, , 1–9.

Toledo, Jon B. et al. 2013. "Clinical and Multimodal Biomarker Correlates of ADNI Neuropathological Findings." *Acta neuropathologica communications* 1(1): 65.

Ulloa, A., S. Plis, E. Erhardt, and V. Calhoun. 2015. "Synthetic Structural Magnetic Resonance Image Generator Improves Deep Learning Prediction of Schizophrenia." In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*,.

Unger, J. W. et al. 1991. "The Amygdala in Alzheimer's Disease: Neuropathology and Alz 50 Immunoreactivity." *Neurobiology of Aging* 12(5): 389–99.

Vieira, Sandra, Walter H.L. Pinaya, and Andrea Mechelli. 2017. "Using Deep Learning to Investigate the Neuroimaging Correlates of Psychiatric and Neurological Disorders: Methods and Applications." *Neuroscience and Biobehavioral Reviews* 74: 58–75.

Visser, P. J. et al. 2002. "Medial Temporal Lobe Atrophy Predicts Alzheimer's Disease in Patients with Minor Cognitive Impairment." *Journal of Neurology Neurosurgery and Psychiatry* 72(4): 491–97.

Walhovd, K B et al. 2010. "Combining MR Imaging, Positron-Emission Tomography, and CSF Biomarkers in the Diagnosis and Prognosis of Alzheimer Disease." *AJNR.American journal of neuroradiology* 31(2): 347–54.

Weiner, Michael W. et al. 2017. "Recent Publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing Progress toward Improved AD Clinical Trials." *Alzheimer's and Dementia* 13(4): e1–85.

Whitwell, Jennifer L et al. 2008. "MRI Patterns of Atrophy Associated with Progression to AD in Amnestic Mild Cognitive Impairment." *Neurology* 70(7): 512–20.

Wilkins, Consuelo H. et al. 2007. "Dementia Undiagnosed in Poor Older Adults with Functional Impairment." *Journal of the American Geriatrics Society* 55(11): 1771–76.

Xie, Saining et al. 2017. "Aggregated Residual Transformations for Deep Neural Networks." In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, , 5987–95.

Yun, Hyuk Jin, Kichang Kwak, and Jong Min Lee. 2015. "Multimodal Discrimination of Alzheimer's Disease Based on Regional Cortical Atrophy and Hypometabolism." *PLoS ONE* 10(6).

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks ArXiv:1311.2901v3 [Cs.CV] 28 Nov 2013." *Computer Vision–ECCV 2014* 8689: 818–33

Zhang, D et al. 2011. "Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment." *NeuroImage* 55(3): 856–67.

FIGURE LEGENDS

Figure 1: A comparison of data demographics and average clinical scores for the studied classes. This study included all subjects in the ADNI repository that passed the minimum selection criterion (minimum follow-up time, conversion or reversion rules) and pre-processing qualitative analysis. Only the baseline scan for each subject was used for all analyses in this study. Clinical scores for diagnosis: MMSE: Mini-Mental State Exam; FAQ: Functional Activities Questionnaire; CDRSB: Clinical Dementia Rating Sum of Boxes; ADAS: Alzheimer's Disease Assessment Scale; RAVLT: Rey Auditory Verbal Learning Test.

Figure 2: A deep residual neural network learning framework is composed of multiple residual blocks that are small stacks of convolutional and batch normalization layers followed by non-linear activation functions such as rectified linear units. In this study, as suggested by the data (Figure 3), we use a model with 3 residual layers for evaluating diagnositic classification performance and progression to AD.

Figure 3: (A) Repeated (n=10) stratified k-fold (k = 5) cross-validation was performed on the pooled cognitively normal (CN) and Alzheimer's Disease (AD) classes to study the effect of adding depth (i.e. adding further convolutional layers or residual blocks) in the implemented framework. Significant improvement in validation accuracy was reported by a model that used 3 residual blocks (D3: depth = 3) as compared to a model that used 2 residual blocks (D2: depth = 2; p = 1.6996e-07) and a model that used 1 residual block (D1: depth = 1; p = 4.5633e-13). Adding another residual block (i.e. depth = 4) did not result in a significant improvement in performance; hence, we've settled on the D3 model and validated it in the several classification/prediction tasks for a consistent comparison. For this specific analysis, all models were run for 100 epochs and used the exact same training and test datasets in each of the cross-validation folds for consistency in performance comparison. (B) The feature spaces at output of the first fully connected layer in the

three surrogate models (for a sample cross-validation fold at the epoch demonstrated by the vertical black line in Figure 3A) were projected onto a two-dimensional space demonstrate additional separation enabled by addition of residual blocks in the 'D3' model as compared to the 'D2' and 'D1' models. The 'Tr' abbrevation corresponds to the training samples whereas 'Te' corresponds to the samples used to test the learnt model.

Figure 4: Six possible binary diagnostic and prognostic classification tasks from the four studied classes were considered. A repeated (n = 10), stratified 5-fold cross-validation procedure was condicted for each of these classification tasks. The ResNet framework was trained independently for each classification task for a maximum of 100 epochs but with an early stopping with a patience level of 20 epochs (20% of the set maximum number of epochs) to prevent overtraining the model. Classification performance was quantified using the accuracy, sensitivity, specificity, and balanced accuracy metrics. Each boxplot shows a spread of the specific reported metric over the 50 cross-validation folds. The first four classification tasks in specific order as in the legend (CN vs. AD, CN vs. pMCI, sMCI vs. AD, and sMCI vs. pMCI) could be considered more clinically relevant and reported a cross-validated median accuracy of 91%, 86%, 86% and 77% respectively, sensitivity of 85%, 79%, 81% and 71% respectively, and specificity of 95%, 92%, 90% and 82% respectively. The performance in the binary classification tasks is comparable or better than previously assessed machine learning architectures while the number of samples is much higher in this specific study. We further explore possible improvements in prediction of progression to AD in the 'Mixed-class prognostic classification' section.

Figure 5: Receiver operating characteristic (ROC) curves were estimated for each of the classification tasks to further evaluate the diagnostic ability of the trained ResNet framework. As expected, the reported area under the curve (AUC) metric follows a similar trend as in Figure 4 thus further adding evidence to the superior performance of the tested architecture for the undertaken analysis.

Figure 6: Mixed-Class Prognosis Classification. A modified form of repeated (n = 10), stratified 5-fold cross-validation procedure was conducted to evaluate the separability of the two MCI sub-classes. Hypothesizing an improvement with an increase in amount of training data provided by other classes (analogous to domain transfer learning), the learner was trained with all datasets from the CN and AD classes (or domains) in addition to the crossvalidation-fold-respective training sMCI/pMCI datasets followed by testing on the crossvalidation-fold-respective testing sMCI/pMCI datasets. (A) and (B) A significant improvement for all studied classification metrics (6% in accuracy, 7% in sensitivity, 5% in specificity and 7% in AUC) was observed for this mixed-class classification task as compared to the standard inter-MCI class classification task (i.e. sMCI vs. pMCI classification task as shown in Figure 4 and bottom left panel in Figure 5). (C) The mixed-class classification task reported a significant performance improvement over the classical SVM model with a p-value of 2.5762e-8.

Figure 7: Multi-class ROC and Classification Projection Analysis. (A) For the multi-class classification, ROC analysis for each class was performed by comparing observations from that class to all other classes (i.e. one vs all comparison). Additionally, micro-averaged and macro-averaged ROC estimates were computed to find singular performance metrics for multi-class classification. Higher AUC was reported by the AD and CN classes followed by the micro-averaged and macro-average cases, while both MCI classes reported lower AUC. (B) and (C) A feature projection analysis was conducted to confirm appropriateness of the learning directionality in the multi-class classification task. In this analysis, the features at the output of the first fully-connected layer in a sample surrogate multi-class model were projected onto a two-dimensional space using the tSNE algorithm. Barring few outliers, the projections of the observations are appropriately ordered by disease severity in terms of diagnostic label (panel B) and clinical scores (panel C). In panel B, the 'Tr' abbrevation in the figure legend corresponds to the training samples whereas 'Te' corresponds to the test samples. In panel C, the following clinical scores were used:- MMSE: Mini-Mental State

Exam, FAQ: Functional Activities Questionnaire, CDRSB: Clinical Dementia Rating Sum of Boxes, ADAS: Alzheimer's Disease Assessment Scale, and RAVLT: Rey Auditory Verbal Learning Test.

Figure 8: (A) The filters for the first convolutional layer were compared in each of the 50 surrogate models (columns 1 to 50 in panel A) as well as the final predictive model (column 51 in panel A) for the modified sMCI vs. pMCI (i.e. mixed-class) classification task. Low pair-wise correlations in the learnt 64 filters in each of these 51 models confirms appropriate learning directionality of the mixed-class framework in the training phase. (B) The mean activations/features at the output of the first convolution layer (normalized and post non-linear activation) were projected back to the brain space to localize activations contributing most to the classification. The 3D activation maps were compared pairwise-across the 51 models pairwise (correlation boxplot on left in panel B); additionally, the 3D activation maps for the 50 surrogate models were compared to the maps from the final predictive model (correlation boxplot on right in panel B). In both cases, a very high median correlation value was observed thus confirming similar features being extracted across the different models. (C) This panel shows the spatial maps of the brain regions corresponding to the peak activations identified by the final predictive model. Hotter (towards red) color imply higher mean activations, while cooler (towards blue) colors imply lower mean activations. The discriminative brain regions at these peak activations were identified in correspondence to the AAL brain atlas.

TABLE 1

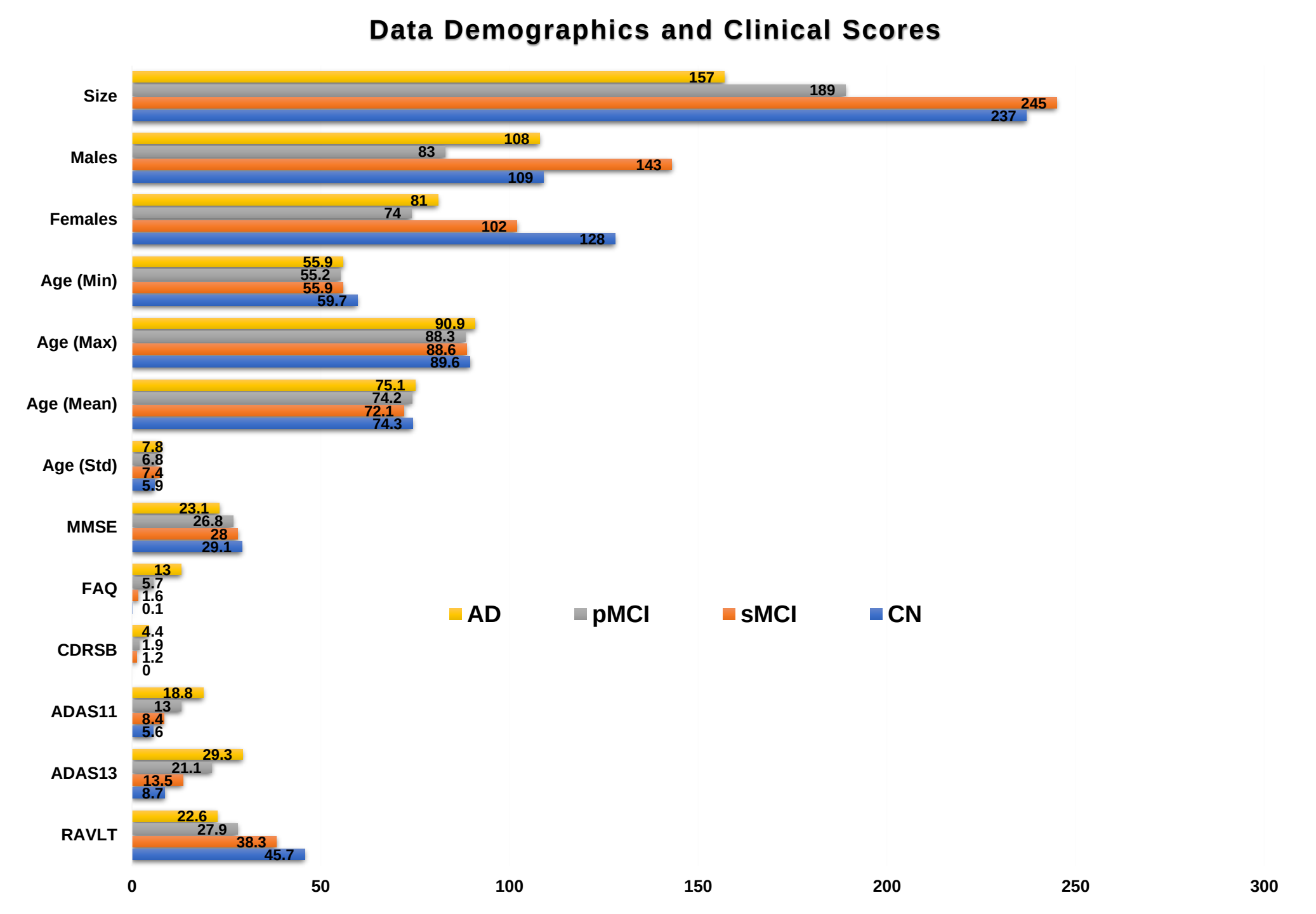
Table 1: Diagnostic/prognostic classification tasks evaluated through the deep ResNet architecture. Standardized 10-repeat, 5-fold (stratified) cross-validation (CV) framework was employed on each of the mentioned tasks except for the mixed-class task (Task TB below) that varied in that the AD and CN classes were also used for training but only the MCI population was used for testing. Classification task TC corresponds to the multi-class classification task where a four-way classification was performed using the same standardized cross-validation procedure.

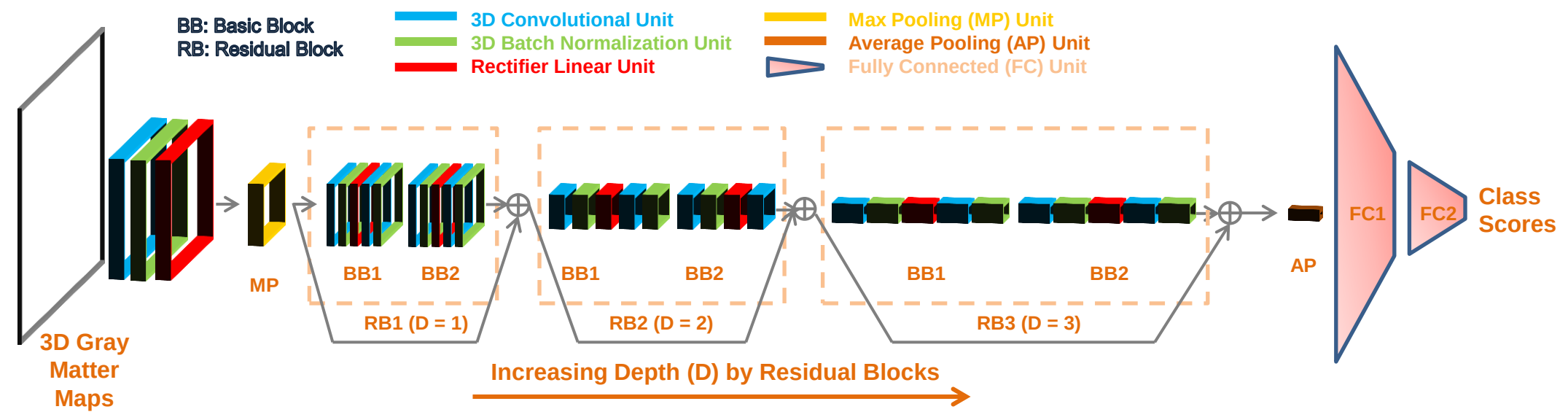
Task	Class 1	Class 2	Class 3	Class 4	5-fold stratified CV (10 repeats)
TA1	CN	AD	-	-	Standard Binary
TA2	CN	pMCI	-	-	Standard Binary
TA3	sMCI	AD	-	-	Standard Binary
TA4	sMCI	pMCI	-	-	Standard Binary
TA5	CN	sMCI	-	-	Standard Binary
TA6	pMCI	AD	-	-	Standard Binary
TB	CN, sMCI	pMCI, AD	-	-	Modified Binary; Split MCI only
TC	CN	sMCI	pMCI	AD	Standard 4-way

TABLE 2

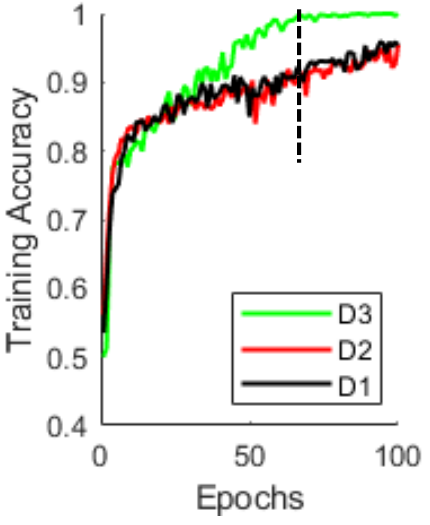
Table 2: Comparison of MCI to AD prediction accuracy using ADNI dataset.

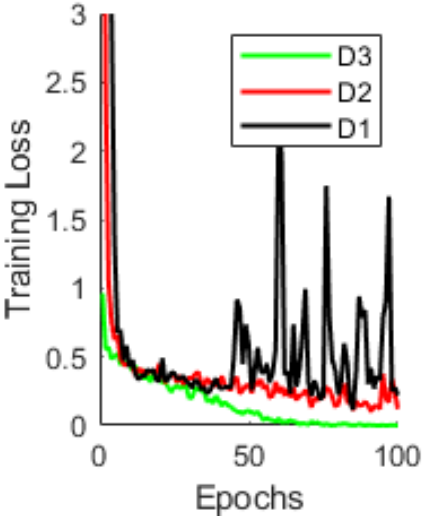
Study	Sample Size	Conversion Period	Architecture	Cross-validation	Accuracy (%)
This work	CN = 237 sMCI = 245 pMCI = 189 AD = 157	36 months	Residual Neural Network	Repeated (n = 10) Stratified 5-Fold	82.7 (MRI)
H. Il Suk et al., 2015	CN = 52 sMCI = 56 pMCI = 43 AD = 51	18 Months	Stacked Auto-Encoder	Repeated (n = 10) 10-Fold	69.3 (MRI) 83.3 (MRI+PET+CSF)
H.-I. Suk, Lee, & Shen, 2015	CN = 52 sMCI = 56 pMCI = 43 AD = 51	18 Months	Deep sparse multi-task learning	Repeated (n = 10) 10-Fold	69.8 (MRI) 74.2 (MRI+PET)
	CN = 229 sMCI = 236 pMCI = 167 AD = 198	18 Months	Deep sparse multi-task learning	Repeated (n = 10) 10-Fold	73.9 (MRI)
Li et al., 2015	CN = 52 sMCI = 56 pMCI = 43 AD = 51	18 Months	Multi-layer perceptron	Repeated (n = 10) 10-Fold	57.4 (MRI+PET+CSF)
H. Il Suk & Shen, 2013b	CN = 52 sMCI = 56 pMCI = 43 AD = 51	18 Months	Stacked Auto-Encoder + Multi-task learning	Repeated (n = 10) 10-Fold	55 (MRI) 75.8 (MRI+PET+CSF +SCORES)
H.-I. Suk, Lee, Shen, & Initiative, 2017	CN = 186 sMCI = 167 pMCI = 226 AD = 226	18 Months	Multi-Output Linear Regression + Deep Convolution Neural Network (CNN)	Repeated (n = 10) 10-Fold	73.28 (MRI+SCORES)
	-- Same --		Joint Linear and Logistic Regression + Deep CNN	Repeated (n = 10) 10-Fold	74.82 (MRI+SCORES)
Shi, Zheng, Li, Zhang, & Ying, 2018	CN = 52 sMCI = 56 pMCI = 43 AD = 51	18 Months	Stacked Deep Polynomial Network	Repeated (n = 5) 10-Fold	78.88 (MRI+PET)
H. Il Suk, Lee, & Shen, 2014	CN = 101 sMCI = 128 pMCI = 76 AD = 93	Unmentioned	Deep Boltzmann Machine	10-Fold	72.42 (MRI) 75.92 (MRI+PET)
Lu, Popuri, Ding, Balachandar, & Beg, 2018	CN = 360 sMCI = 409 pMCI = 217 AD = 238	0 to 36 Months	Multiscale Deep Neural Network	10-Fold	75.44 (MRI) 82.93 (MRI+PET)

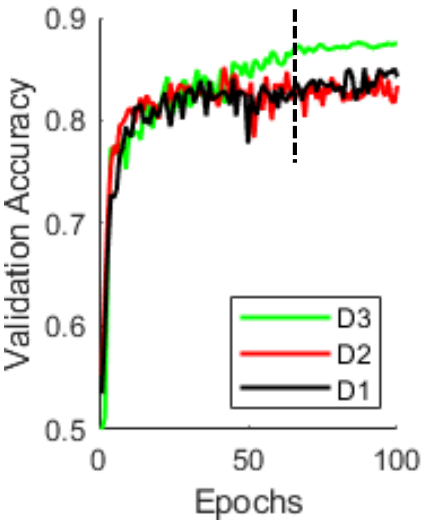

Data Demographics and Clinical Scores
Size
157
189
245
237
Males
108
83
143
109
Females
81
74
102
128
Age (Min)
55.9
55.2
55.9
59.7
Age (Max)
90.9
88.3
88.6
89.6
Age (Mean)
75.1
74.2
72.1
74.3
Age (Std)
7.8
6.8
7.4
5.9
MMSE
23.1
26.8
28
29.1
FAQ
13
5.7
1.6
0.1
CDRSB
4.4
1.9
1.2
0
ADAS11
18.8
13
8.4
5.6
ADAS13
29.3
21.1
13.5
8.7
RAVLT
22.6
27.9
38.3
45.7
AD
pMCI
sMCI
CN
0
50
100
150
200
250
300


BB: Basic Block
RB: Residual Block
3D Convolutional Unit
3D Batch Normalization Unit
Rectifier Linear Unit
Max Pooling (MP) Unit
Average Pooling (AP) Unit
Fully Connected (FC) Unit
3D Gray Matter Maps
MP
BB1
BB2
RB1 (D = 1)
BB1
BB2
RB2 (D = 2)
BB1
BB2
RB3 (D = 3)
AP
FC1
FC2
Class Scores
Increasing Depth (D) by Residual Blocks

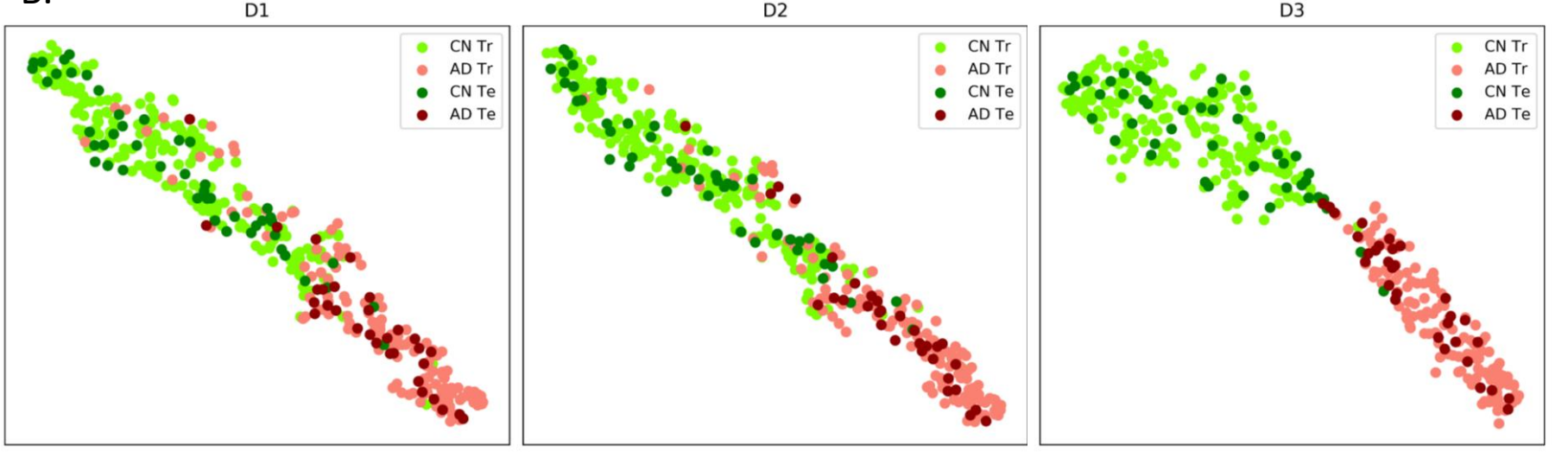
A.

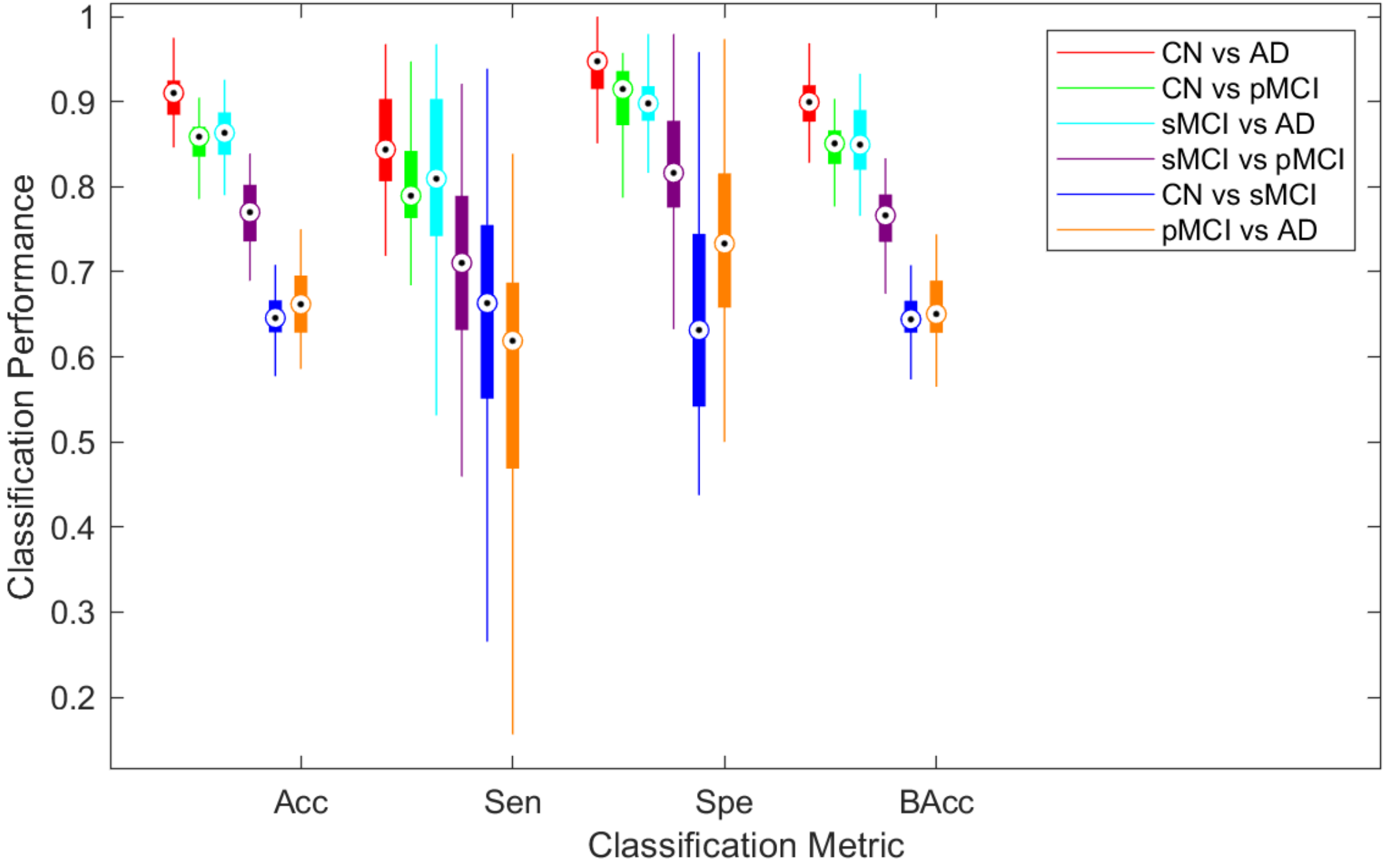

Training Accuracy
Epochs
D3
D2
D1

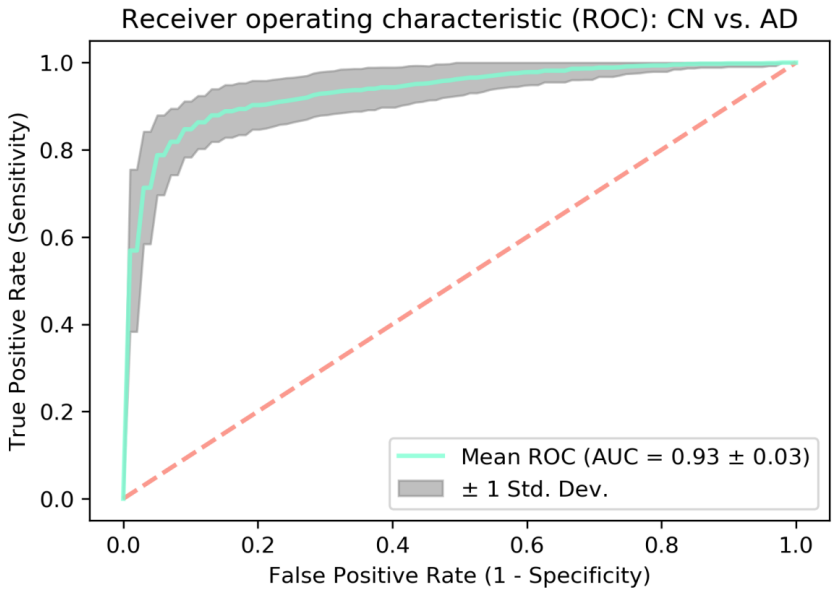

Training Loss
Epochs
D3
D2
D1

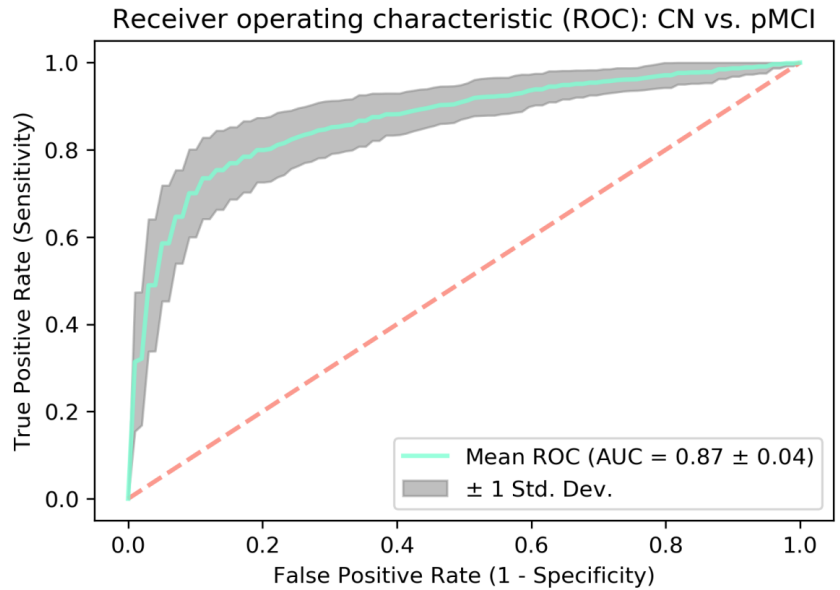

Validation Accuracy
Epochs
D3
D2
D1

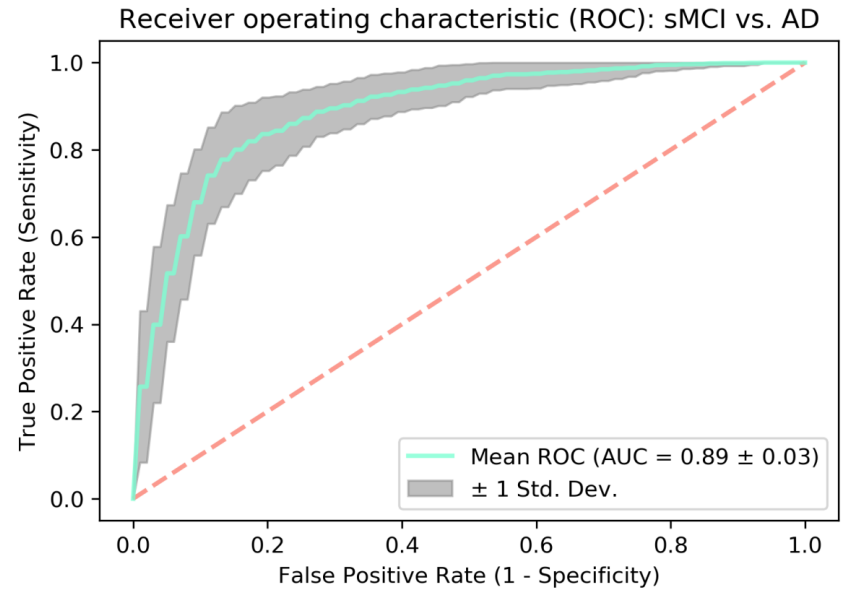
B.

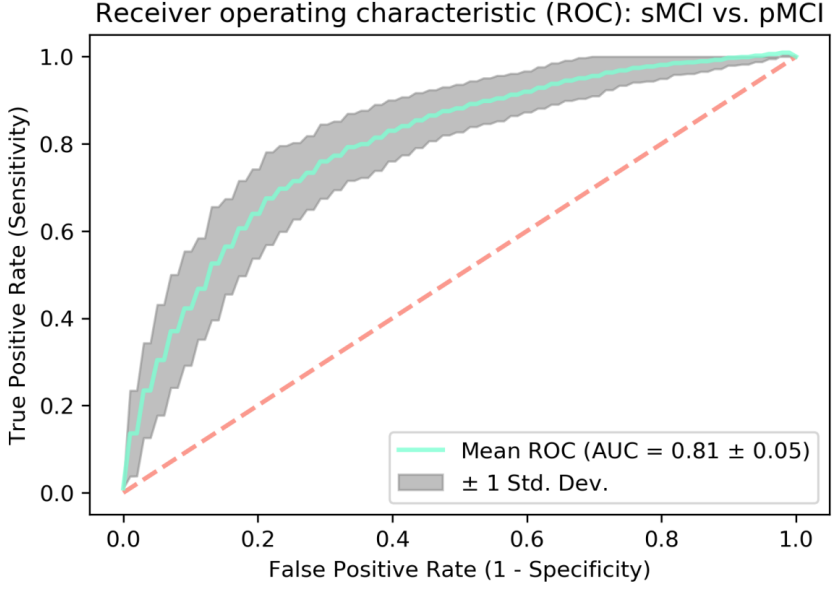

D1
D2
D3
CN Tr
AD Tr
CN Te
AD Te

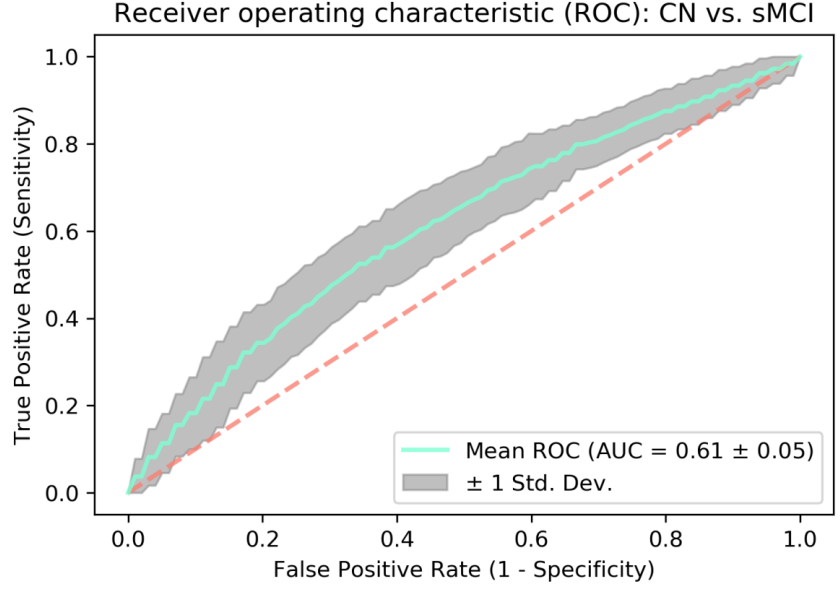

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Classification Performance
Acc
Sen
Spe
BAcc
Classification Metric
CN vs AD
CN vs pMCI
sMCI vs AD
sMCI vs pMCI
CN vs sMCI
pMCI vs AD

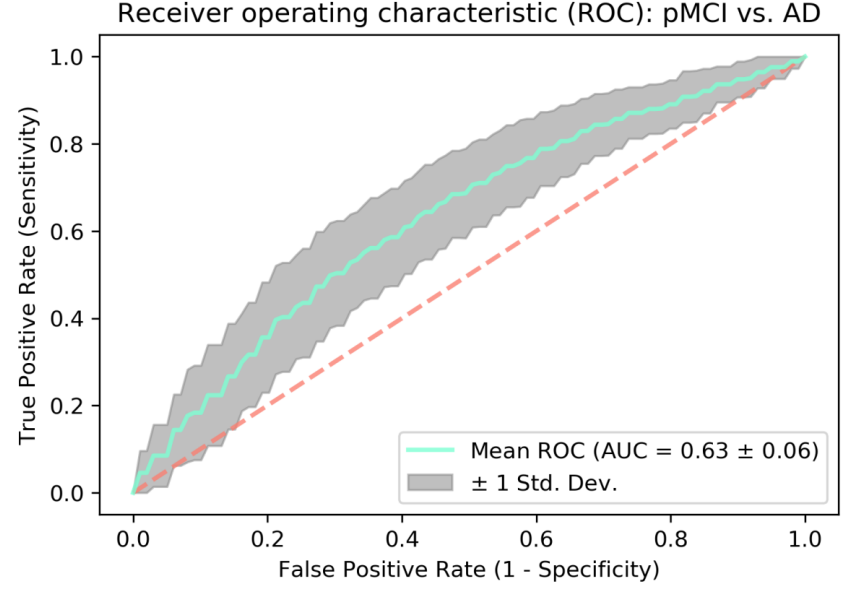

Receiver operating characteristic (ROC): CN vs. AD
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.93 ± 0.03)
± 1 Std. Dev.

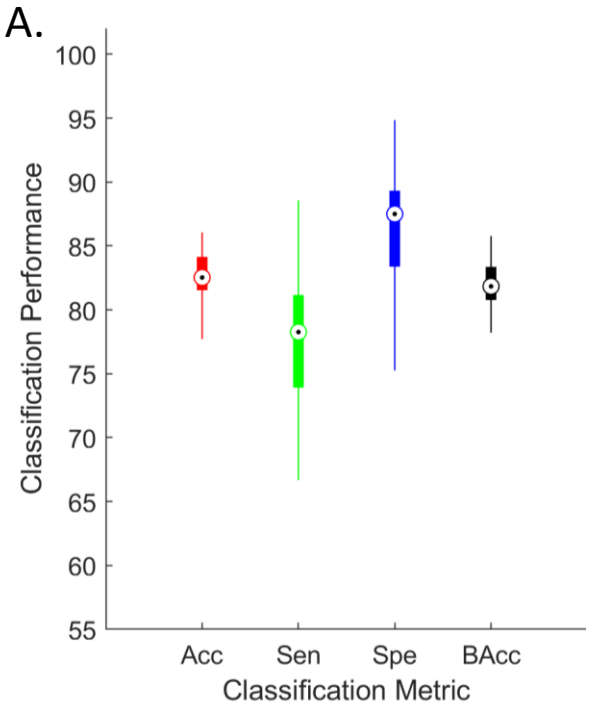

Receiver operating characteristic (ROC): CN vs. pMCI
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.87 ± 0.04)
± 1 Std. Dev.

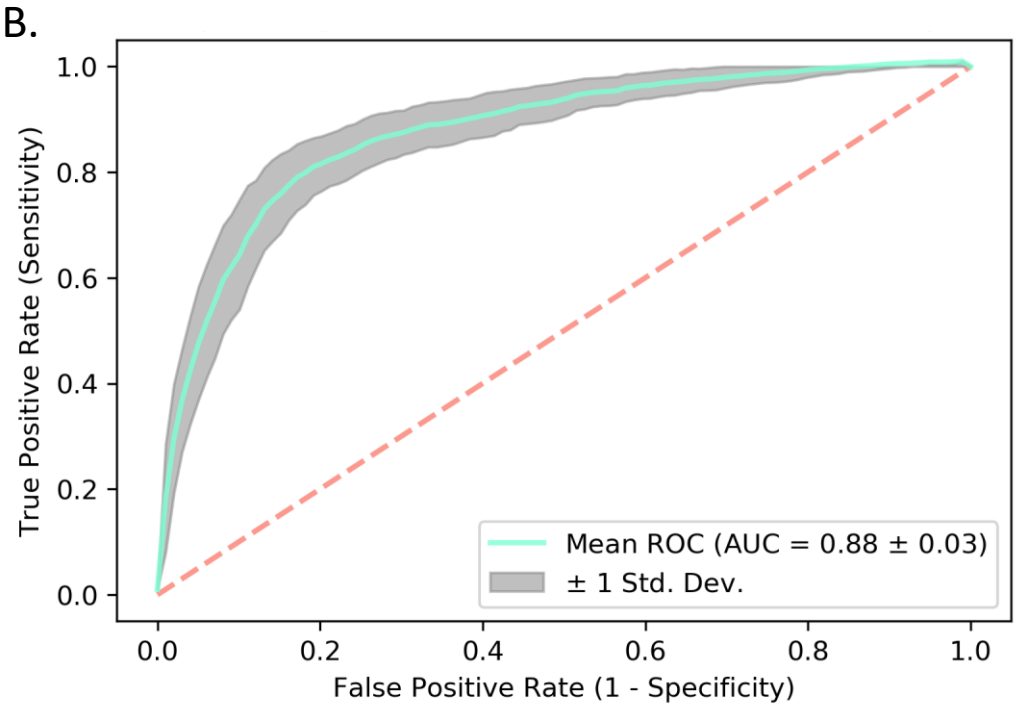

Receiver operating characteristic (ROC): sMCI vs. AD
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.89 ± 0.03)
± 1 Std. Dev.

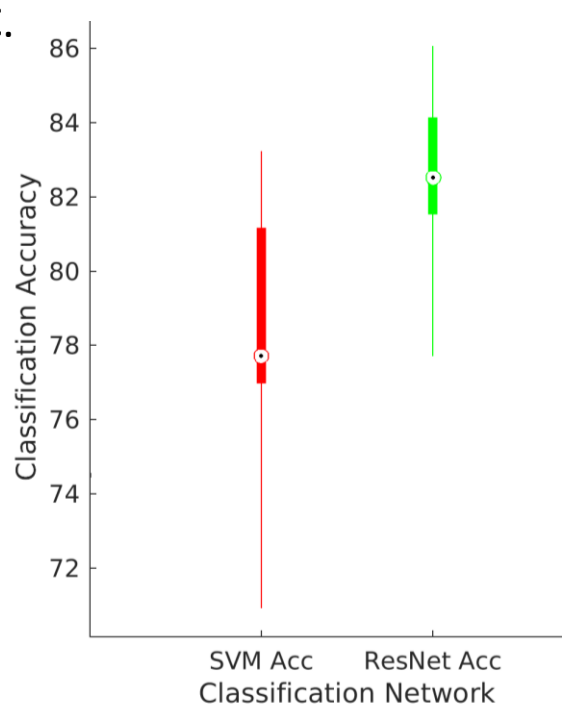

Receiver operating characteristic (ROC): sMCI vs. pMCI
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.81 ± 0.05)
± 1 Std. Dev.

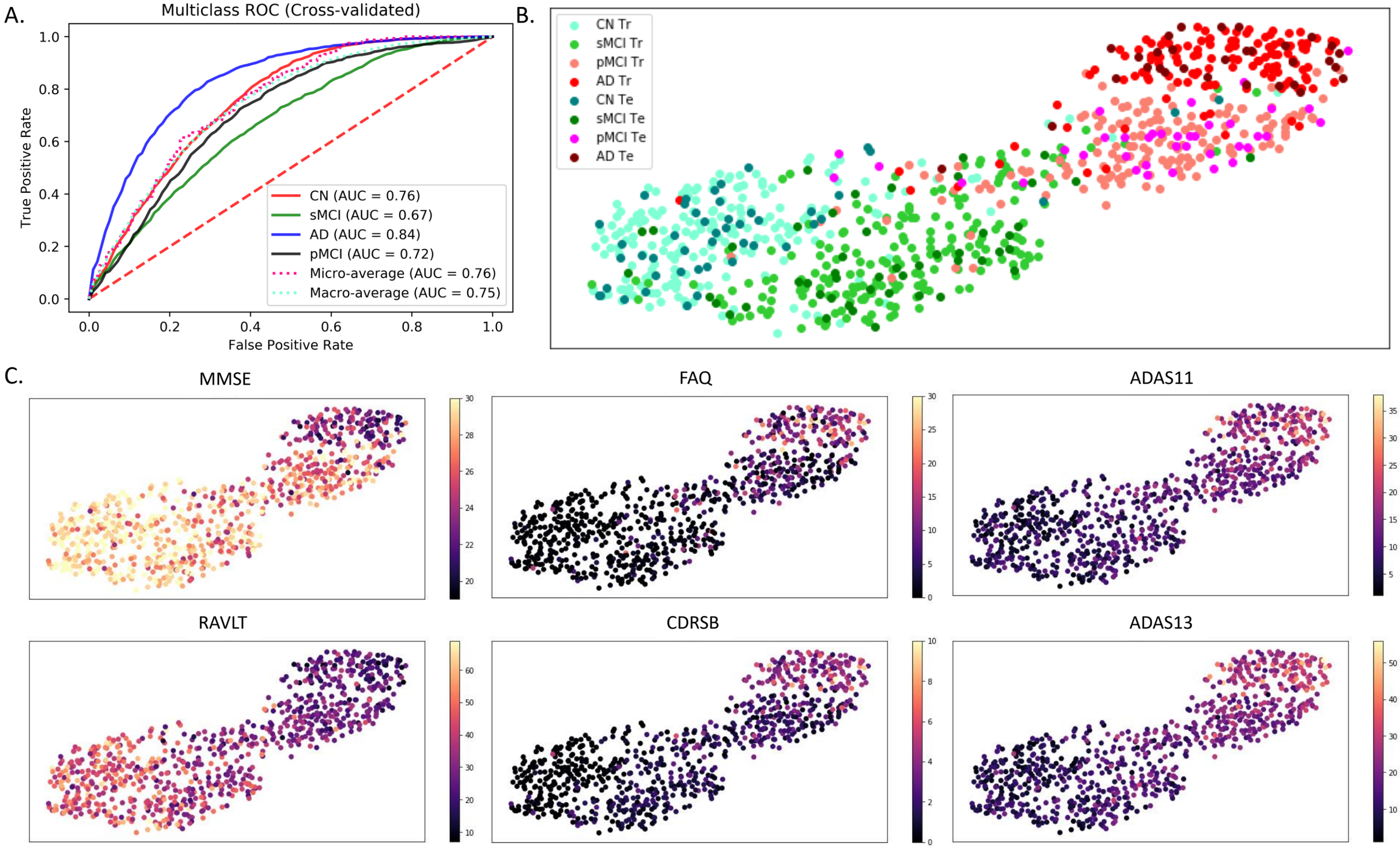

Receiver operating characteristic (ROC): CN vs. sMCI
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.61 ± 0.05)
± 1 Std. Dev.


Receiver operating characteristic (ROC): pMCI vs. AD
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.63 ± 0.06)
± 1 Std. Dev.


A.
Classification Performance
100
95
90
85
80
75
70
65
60
55
Acc
Sen
Spe
BAcc
Classification Metric


B.
True Positive Rate (Sensitivity)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (1 - Specificity)
Mean ROC (AUC = 0.88 ± 0.03)
± 1 Std. Dev.


C.
Classification Accuracy
86
84
82
80
78
76
74
72
SVM Acc
ResNet Acc
Classification Network


A.
Multiclass ROC (Cross-validated)
True Positive Rate
False Positive Rate
CN (AUC = 0.76)
sMCI (AUC = 0.67)
AD (AUC = 0.84)
pMCI (AUC = 0.72)
Micro-average (AUC = 0.76)
Macro-average (AUC = 0.75)
B.
CN Tr
sMCI Tr
pMCI Tr
AD Tr
CN Te
sMCI Te
pMCI Te
AD Te
C.
MMSE
FAQ
ADAS11
RAVLT
CDRSB
ADAS13