Multi-task Learning and Ensemble Approach to Predict Cognitive Scores for Patients with Alzheimer's Disease

†Daren Ma[1], **†Christabelle Pabalan**[2], **Akanksha .**[2], **Yannet Interian**[2‡], **Ashish Raj**[1‡]

[1]Department of Radiology, University of California, San Francisco; [2]University of San Francisco

***For correspondence:**
daren.ma;ashish.raj@ucsf.edu (1);
yinterian;cpabalan@dons.usf.edu (2)

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Present address: [§]Department of Radiology, University of California, San Francisco

Abstract The objective of this research was to explore the efficacy of integrating **1)** using multi-task learning, **2)** neuroimaging data and common risk factors to predict the ADAS-cog 11 score in Alzheimer's patients. We applied heavy augmentations to magnetic resonance imaging (MRI) scans of 798 participants ranging between 0 and 96 months from the initial diagnosis. This enabled us to exploit the benefits of the U-Net architecture, a structure historically known to perform well on medical imaging tasks using limited data with heavy image augmentations. The multi-task model simultaneously performed segmentation of white matter, gray matter, and cerebrospinal fluid on the MRI input and regression to predict the ADAS-cog 11 score. There is a body of literature highlighting the independent relationships between gray and white matter volumes and dementia severity; the trajectory was to explore if the multi-task model structure for these interrelated tasks would boost performance for each individual task. The final model integrates the deep learning results with a Gradient Boosting model trained on demographic data. A considerable performance improvement from the initial multi-task U-Net on the ADAS-cog 11 prediction task was achieved across our experiments.

1 Introduction

Early detection of Alzheimer's disease can dramatically reduce the burden of the disease by enabling healthcare professionals to recommend lifestyle changes that can slow the progression of the disease and allowing patients to plan ahead for the future. The emergence of deep learning techniques in the medical field has seen rapid success in recent years and has been rising to a prominent position in medical imaging and diagnosis. With this recent explosion, there is an emerging body of literature that revolves around improving early Alzheimer's diagnosis using machine learning methods and neuroimaging data. However, there are "several risk factors in subjects genetically, demographically, and lifestyle [that] are more important than some medical history risk factors" (Mahyoub et al., 2018). Team Zandifar's work ***Zandifar et al.*** (***2020***) illustrated that "Combining cognitive scores and MRI biomarkers yield accurate prediction years before onset of dementia ... may be helpful in selecting patients that would most benefit from lifestyle changes, and eventually early treatments that would slow cognitive decline and delay the onset of dementia". This work opened the path of our current and future research, and perfectly emphasized the potential of this model in helping patients in the clinical service process.

Our paper presents an integrated method to perform Alzheimer's prognosis using neuroimaging data along with tabular data composed of common risk factors based on voxel-level MRI images specifically.

Our pipeline has three phases: a segmentation model to segment MRI scans into white matter, gray matter and cerebrospinal fluid, a multi-task model that performs the same segmentation task with an additional regression task to predict cognitive scores, and an ensemble model that also performs both segmentation and regression. **Figure 4** outlines the different stages of modeling, expanding on the tasks and the inputs fed into every model.

The objective of this study was to investigate the extent to which we could predict a cognitive score in Alzheimer's patients using machine learning techniques. We began by developing a U-Net that specializes in segmenting the MRI scans into white matter, gray matter, and cerebrospinal fluid to create a baseline. Afterward, we added a linear layer at the bottom of the U-Net, enabling the U-Net to output a number representing the ADAS11 cognitive score. A model containing a simultaneous segmentation and regression task is an instance of multi-task learning; according to Wikipedia, this form of learning is "a sub field of machine learning in which multiple learning tasks are solved at the same time while exploiting commonalities and differences across tasks [...] using the domain information contained in the training signals of related tasks". In other words, "what is learned for each task can help other tasks be learned better".

Moreover, in the work ***Stout et al.*** (***1996***) of Stout et al., the authors note that "quantitative magnetic resonance methods provided strong evidence that cortical gray matter volume, which may reflect atrophy, and abnormal white matter volume are independently related to dementia severity in probable Alzheimer disease: lower gray matter and higher abnormal white matter volumes are associated with more severe dementia". Consequently, we hypothesized that a multi-task model consisting of segmentation would allow the U-Net to understand the context of the ratio of white and gray matter to potentially boost performance on the cognitive score prediction.

2 Methods and Materials

2.1 Dataset

Alzheimer's Disease usually observed in elderly individuals can be diagnosed using the ADNI dataset, which contains details of MRI scans for 843 subjects with scanner intensity fields ranging from 1.5 T to 3 T. It is observed that patients with lower cognitive abilities including thinking and memory loss are associated with mild cognitive impairment (MCI). They also have a high risk of transforming to AD or any other types of dementia and are grouped separately from AD.

Training data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD, with an overall goal to validate biomarkers for use in clinical treatment trials for patients with Alzheimer's Disease (AD). The study was approved by the Institutional Review Boards at each ADNI site (see full list here: http://adni.loni.usc.edu). Informed consent was obtained from all subjects prior to enrollment. All methods were carried out in accordance with relevant guidelines and regulation. In the present study, we included subjects from the ADNI dataset defined as cognitively normal (CN) or as patients with an MCI or AD diagnosis.

In this work, we selected the subjects' ADAS-Cog scores to be our major target. The Alzheimer's Disease Assessment Scale-cognition sub-scale (ADAS-Cog) (Rosen et al. 1984) is the most widely used general cognitive measure in clinical trials of AD (Connor and Sabbagh 2008; Ihl et al. 2012; Rozzini et al. 2007).

The secondary training target is the brain segmentation of the MRI images. We performed 3D semantic image segmentation task of brain MRIs into White Matter, Grey Matter, and Cerebrospinal fluid. The ground truth was generated by the FAST pipeline **?**, a computational tool that segments a 3D image of the brain into different tissue types, whilst also correcting for spatial intensity variations. It is robust and reliable, compared to most finite mixture model-based methods, which are sensitive to noise.

For training-testing separation, we put ADNI Stage-3 data aside as the separate testing set, which is not included during the training process. We took ADNI 1, 2 datasets as our training set,

and applied cross-validation to tune the model. In the following table, we report the demographic distribution of the experimented population.

Characteristic	Set 1 Training Set	Set 2 Validation Set	Set 3 Set-aside Training Set
Gender (F/M)	246/284	95/134	40/44
Marriage (Married/Divorced/Never Married)	398/43/18	172/21/11	69/6/2
Age (mean, range)	73.2 (55-91.4)	72.8 (55-89.6)	73.5 (55-86.1)
Years of Education (mean, range)	16.2 (8-20)	15.7 (6-20)	16.1 (8-20)
MMSE (mean, range)	26.1 (4-30)	26.1 (7-30)	25.5 (0-30)
ADAS-Cog-11 (mean, range)	12.76 (0-63)	12.34 (0-58)	14.30 (2-69)

Table 1. Demographic distribution among the data sets

2.2 Training Target

First introduced in the 1980s***Rosen et al.*** (***1984***), ADAS-Cog is widely accepted as a gold standard for assessing the efficacy of anti-dementia treatments***Kueper et al.*** (***2018***). We chose ADAS-Cog-11 as our primary research goal for the following reasons. First, it's predictive in the subjects' current cognition status as well as dementia development. Second, we would like to build a Deep Learning architecture that solves not only the traditional segmentation task, which can be also done by FreeSurfer, but also the score prediction task to help clinical service in multiple ways. Segmentation task is mainly used as a regularization tool of this topic.

We selected the baseline ADAS-Cog-11 scores of the subjects, namely the cognitive scores sampled when they went in to the clinics for the first time. In our future study, we aim at solving the problem of predicting the changes in the cognitive scores after a time period no shorter than 6 months.

2.3 Data Preprocessing

For the tabular demographic data, we first filtered out the regional biomarker features from the ADNI dataset, and then conducted one-hot encoding for the categorical features including marriage status and gender. We went through this data cleaning process in order to make sure that there are only survey-based data to be introduced to the tabular regressor.

For the ADNI Image data, we performed data augmentation to artificially generate new training images by applying transformations to existing images; these transformations consist of operations such as zooms, flips, shifts, deformations, etc. These techniques are particularly beneficial in the medical imaging space where there's a scarcity of de-identified, readily available training instances. Particularly for medical imaging, it's critical to be sensitive to the problem domain and ensure we're not generating unrealistic synthetic examples through the use of deformation transformations. However, for brain imaging, these deformations can "play an important role ... as they are able to preserve topology and generate biologically plausible deformations" ***Nalepa et al.*** (***2019***)

The finalized pipeline of transformations applied to the MRI training dataset included affine image transformations (random crop to 128x128x128), elastic transformation to introduce shape variation to the data, and normalization to optimize learning performance. For the validation set, we performed a center crop to 128x128x128 and normalization. Since the validation images serve as the testing ground truth, we use center crop to maintain the same domain as the training set yet capture the same consistent region.

3 Multi-task Network Architecture

3.1 Baseline CNN Image Model

As the first step of our study, we trained a Convolutional Neural Network (CNN) model on the image data. This basic CNN network was composed of the following layers: first convolutional layer,

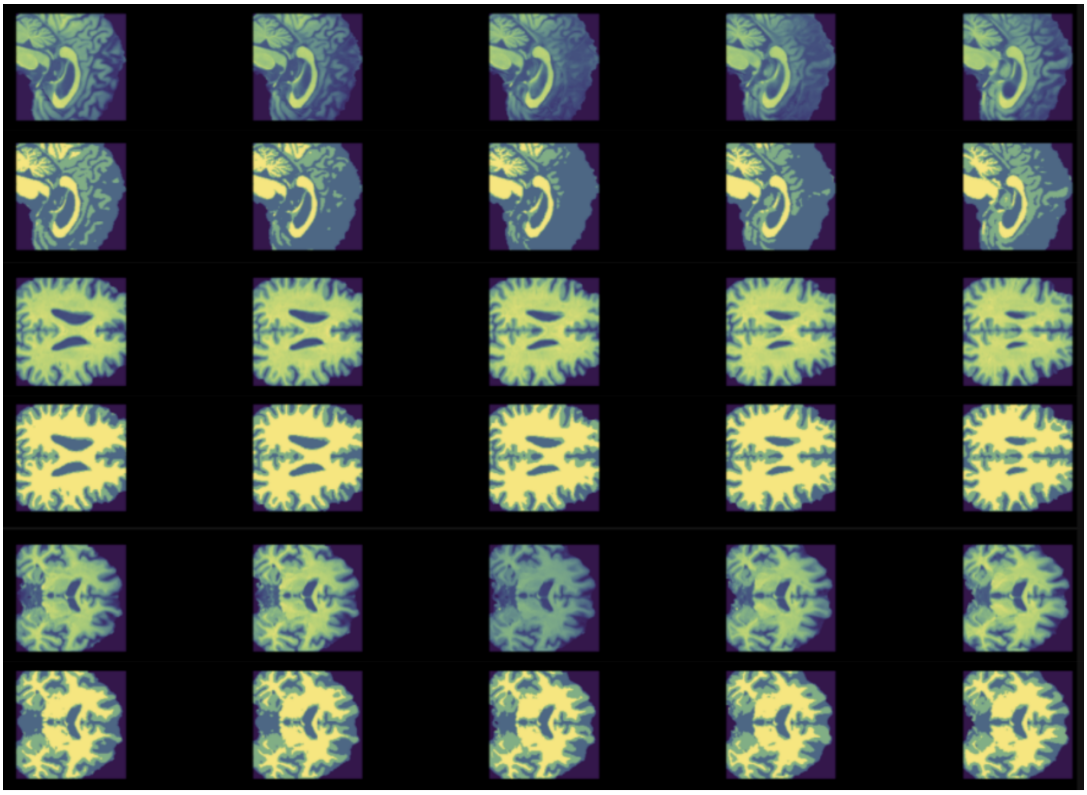


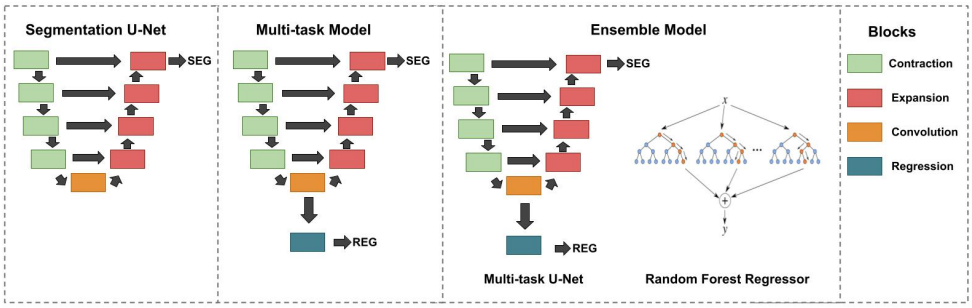
Figure 1. MRI Scans After Data Augmentations

pooling layer, second convolutional layer, pooling layer, a 2-layer fully connected Neural Network, and the output layer. During the first approach, we selected the convolutional kernel size to be $2 \times 2 \times 2$. We adapted the global average pooling layer introduced by popular methods such as ResNet. The output layer utilized a Linear Regression method with a Min-Squared Error (MSE) loss for the ADAS-Cog-11 score. By adapting the ideas of the work by Folego et al. ***Zhang et al.*** (***2001***), we enlarged the convolutional kernel size to $3 \times 3 \times 3$, and applied a dropout parameter of 0.5 for regularization.

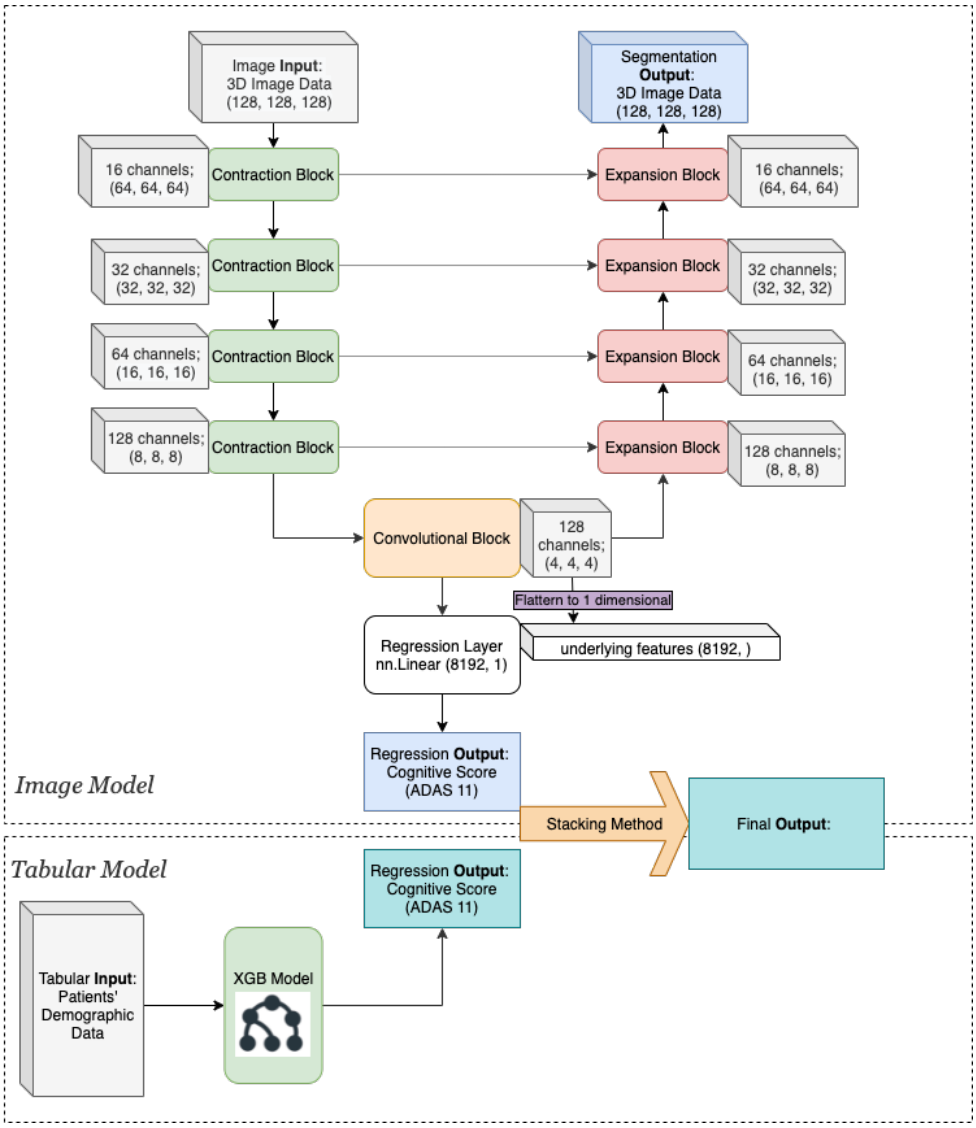
3.2 Image Model Architecture

The U-Net architecture performs effectively for medical imaging tasks with limited data and the presence of heavy data augmentations. The structure consists of two integral components to maximize segmentation performance. The first is a contracting path where the input size is downsampled but the depth of the feature map increases. This allows us to 1) enhance training speed due to the decreasing number of computation and 2) introduce greater generalizability by integrating a small location invariance. However, to regain that precise localization necessary for accurate segmentation, the second component of the architecture is an expanding path where the feature map is up-sampled. During this reconstruction phase, the architecture also consists of skip connections where contextual information in the corresponding cropped feature map from the contracting path is concatenated to the expansion layers.

Our multi-task architecture is illustrated in **Figure 1** It consists of the assumed contracting and expanding paths, each with four resolution steps. In the contracting path, each layer contains two $3 * 3 * 3$ convolutions (with a padding of 1), followed by a rectified linear unit (ReLu), and then a $2 * 2 * 2$ max-pooling operation (with strides of two in each dimension). In the synthesis path, every layer consists of a $2 * 2 * 2$ up-convolution (with strides of two in each dimension), followed by two $3 * 3 * 3$ convolutions and a ReLu operation. Shortcut connections from layers of equal



(a) Modeling Pipeline



(b) 3D U-Net architecture. The number of channels is detonated above each feature map (represented as the blue boxes)

Figure 2. Model Structure

resolution in the contracting path provide the essential high-resolution features to the synthesis path, as noted previously. The last layer of our architecture consists of a $1 * 1 * 1$ convolution that reduces the number of output channels to the number of labels, which is 3 labels for our segmentation task.

The loss function for the ensemble model is that: $Loss = w1 \cdot Loss_U + w2 \cdot Loss_H$, where $Loss_U$ is the ADAS-Cog-11 prediction loss from the 3D U-Net model, and $Loss_H$ is the MSE Loss from the tabular HGB model.

1. **Convolutional Layers:** filters are applied to the original image based on the number of kernels and the size of the kernels
2. **Max-pooling Layers:** returns the maximum value in a certain filter region to reduce the dimensionality of the network
3. **Batch Normalization:** Standardizes the inputs to a layer for each mini-batch. This stabilizes the learning process and dramatically reduces the number of training epochs required to train deep networks.
4. **ReLu:** $f(x) = max(0, x)$, introduces non-linearity to the model, helping the network learn complex patterns in the data and provides output from 0 to infinity

3.3 Tabular Regressor

We selected Histogram-based Gradient Boosting regressor (HGB Regressor) from scikit-learn(citation here) for the tabular demographic data.

We applied cross-validation to train the model on ADNI stage 12 demographic data, with a 0.7/0.3 split ratio. We selected the Poisson loss function, and fixed the max depth of decision trees to be 6 layers. Based on grid search training method, the optimal lambda for the L2 regularization method was 2.

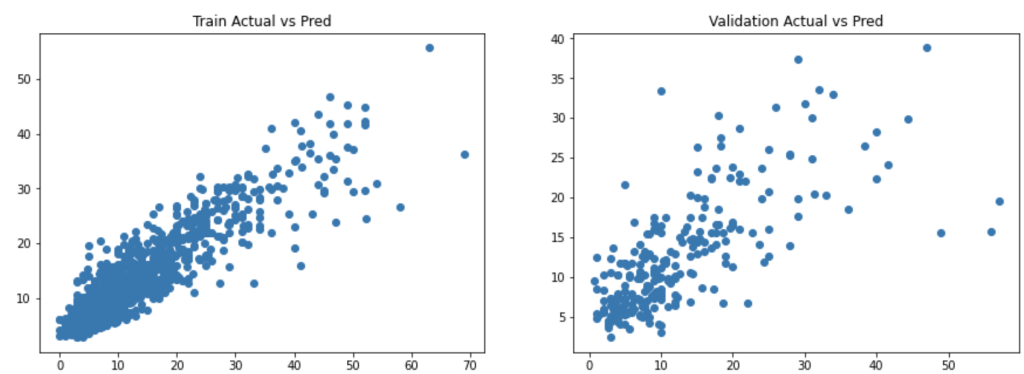


Figure 3. The scatter plot of prediction vs. real target

3.4 Multi-task Ensemble Model

In the subsequent experiment, we demonstrate the efficacy of adding demographic data and potential risk factors of Alzheimer's into our model. The challenge of merging tabular data and image data is the nature of the image data containing a significantly larger number of inputs (the number of voxels) in opposition with tabular data where a single feature is commonly represented with a single value. As a result, the tabular features can easily get lost in translation during the event of parameter optimization embedded in a deep learning architecture. To account for this, we used the gradient boosting regressor mentioned in section 3.3 to predict the ADAS11 cognitive score based on the tabular data, and implemented a multi-task, 3D U-Net for the image data to perform segmentation and regression jointly. We integrated the regression tasks by taking the weighted

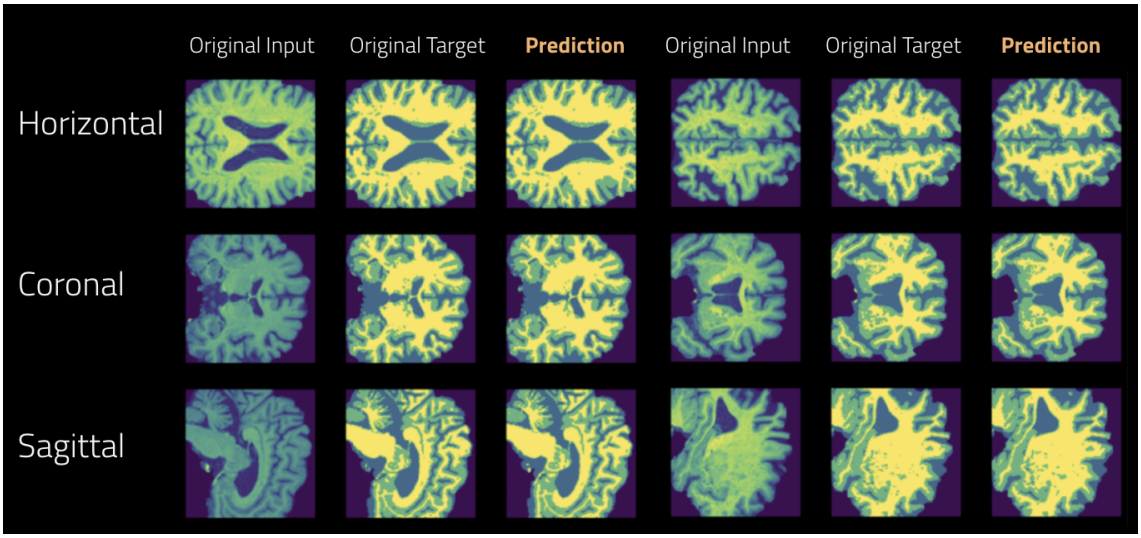


Figure 4. This figure plots the raw MRI image data, Segmentation ground truth, and our Segmentation results in slices of 3 dimensions.

sum of both predictions using their R-squared scores as the indicator of the weights of the corresponding models.

We have also considered other stacking approaches. We added a multi-input layer to the U-Net model which brings in the tabular data, and made a residual combination in the forward function. This got us comparable results with our best model. Another approach was to train a Gradient Boosting Algorithm along with the U-Net and update the loss function of the GB model correspondingly.

4 Experimental Results

4.1 Segmentation Performance

As shown in **Figure 2**, the U-Net architecture was able to generate segmentation results close to ground truth. In terms of metrics, the segmentation task of our best model achieved a validation dice coefficient of 0.9978, 0.9635, 0.9513, and 0.9725 for background, gray matter, white matter and cerebrospinal fluid, respectively.

The Dice similarity coefficient (DSC) was used as a statistical validation metric to evaluate the performance of both the reproducibility of manual segmentations and the spatial overlap accuracy of automated probabilistic fractional segmentation of MR images, illustrated on two clinical examples.

4.2 Regression Performance

Model	MSE	$DiceScore$	$Cross-Validation\ R^2 Range$	$Testing\ R^2$
Gradient Boosting Regressor (Tabular)	44.92	-	0.56 - 0.63	0.51
Single-Task CNN Model	42.14	-	0.33 - 0.52	0.47
Multi-Task U-Net Model	51.58	0.9793	0.41 - 0.69	0.57
Ensemble Model using MLP	38.30	0.9812	0.65 - 0.77	0.64
R-squared Ensemble Model	**27.11**	**0.9812**	**0.73 - 0.78**	**0.67**

Table 2. Model Results for ADAS-Cog 11 prediction

Table 1 shows the experimental results on the original AD task. Note that a HGB regressor

alone produces an R-squared over 0.5, showing that the tabular data are decent indicators for the cognitive score. The Single-task model achieved a mediocre R squared comparing with the tabular data. Its performance on the set-aside testing set was even worse than the basic tabular model, implying that a more complicated architect is probably necessary. The Multi-task model does the ADAS-Cog score regression based on its accurate segmentation, and achieves lower training loss and higher R-squared than both the tabular model and the Single-task model, which proves our assumption back in the introduction. So far we have got great performance in the regression task, but the ensemble method significantly improved the results.

Our first stacking attempt was to design a 2-layer MLP which took in both the tabular demographic data and the U-Net bottom-layer latent vectors. This gave us an MSE around 38.0 plus an R-squared of 0.64. Also, we trained a Convolutional Neural Network model for the prediction task alone using ADNI Image data. The results are reported in Table 1.

Then, inspired by the idea of ResNet ***He et al.*** (***2015***), we created a special R-squared Ensemble model, assuming that the R-squared score of each separate model measures the amount of interpretability it adds to the whole task. As shown in the last row, our R-squared Ensemble model got the highest metrics over all the experiments.

These results indicate that the underlying features captured from the segmentation task was predictive for the AD task at ADNI dataset.

4.3 Hyper Parameter Tuning

Training Parameters	Segmentation U-Net	Multi-task U-Net	Ensemble Model
Batch size	10	12	12
Epochs	80	62	45
Weight Decay	1e-5	1e-6	1e-6
Optimizer	Adam	AdamW	AdamW

Table 3. Training Parameters for Each Model

Figure 5 shows our hyper parameter conducting with the 10-fold cross-validation. First we fixed the dropout levels of the model at 1e-6, 1e-7, and 1e-8, and printed out the validation MSE loss versus the training epochs. The ensemvble model reached its optimal validation results on 45 training epochs in all three dropout levels. Then, we fixed the training epoch to be 45, and found that 1e-6 is the best dropout regularization parameter for our U-Net based multi-task model.

Table 2 displays the training parameters used during training for the U-Net, segmentation model, the multi-task model, and the ensemble model, respectively.

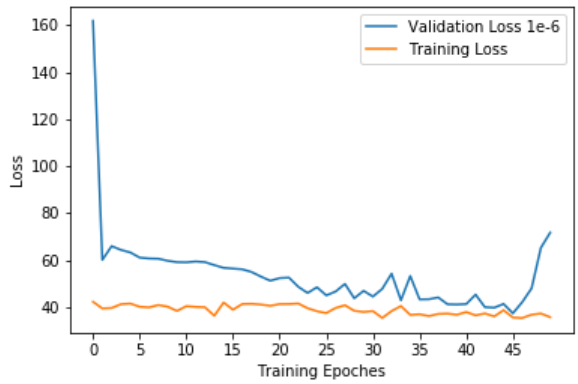
5 Discussions

5.1 Comparison to existing prediction methods

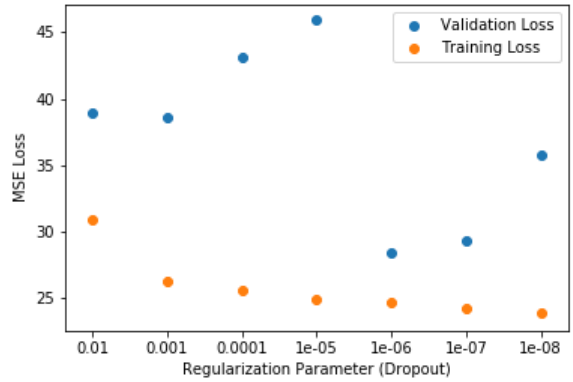
Our model results show that the U-Net architecture is able to predict the cognitive scores directly from the subjects' MRI Image data.

Back in 2010, Stonnington et al applied Relevance vector regression (RVR) to the prediction of clinical ratings based on the MR structural images from two independent data bases, and included ADAS-Cog score as one of their major training goals ***Stonnington et al.*** (***2010***). RVR is a sparse kernel method formulated in a Bayesian framework, and the authors got an R-squared of 0.58 for the ADAS-Cog task. Considering this was a work 10 years ago, this is a very high performance. RVR models required more pre-processing steps than the integrated U-Net model, including a Jacobian transformation, which can slow down the training process. (Talk about their cross validations.)

Zhang et al. (***2017***) is a similar previous work that also predicted cognitive scores using a multi-task deep learning model. The authors designed a multi-task dictionary learning CNN model which



(a) Model Loss vs Training Epochs with weight decay=1e-6



(b) Model Loss vs Dropout level

Figure 5. Hyper Parameter Tuning

was pre-trained on the rich ImageNet data set, and they transferred learning it to take the baseline hippocampal, ventricular and cortical thickness surface features as the model input. The model produced superior results in predicting both MMSE and ADAS-Cog scores. Their best model achieved a weighted R of 0.862±0.045 on the ADNI ROI data set, equivalent to an R-squared measure of 0.743. However, given the large pre-training set and longer training time, this work didn't perform much better than our final model. What's more, their results were reported based on a 10-fold cross-validation only without a set-aside testing set.

Choi and Jin (***2018***) This work utilized a 3D CNN model to predict the subjects' conversion from MCI to AD using their PET scans. The experimental results showed that the CNN-based biomarker strongly correlated with future cognitive decline. Accuracy of prediction (84.2 %) for conversion to AD in MCI patients, which outperformed conventional feature-based quantification approaches, expressed the feasibility of deep learning as a practical tool for developing predictive neuroimaging biomarker.

Popescu et al. (***2021***) developed a 3D U-Net model to predict chronological age at the local level in healthy people, by incorporating voxelwise information using deep learning . Their local brain-age model on healthy participants combined from the OASIS3 (n = 128), AIBL (n = 83) and Wayne State (n = 200) datasets. And the corresponding R scores were 0.47, 0.78, and 0.72.

The major outstanding value of our work is that we specifically used MRI images alone to generate the predictions of ADAS-Cog-11, a metric traditionally being collected in laboratory, and achieved reasonable modeling results. Our R-squared results are higher than or at least similar to these previous work using more complicated U-Net structures, showing that the simplified R2 stacking has strong potential in getting comparable results while saving computational power.

5.2 Next Steps

Given the outstanding performance of a multi-task Deep Learning model on a prediction problem combining image MRIs and demographic data, our team reached out to other research groups for the PPMI dataset on the Parkinson Disease.

References

Balsis S, Benge JF, Lowe DA, Geraci L, Doody RS. How Do Scores on the ADAS-Cog, MMSE, and CDR-SOB Correspond? Clin Neuropsychol. 2015; 29(7):1002–1009.

Choi H, Jin KH. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behavioural Brain Research. 2018; 344:103–109. https://www.sciencedirect.com/science/article/pii/S0166432818301013, doi: https://doi.org/10.1016/j.bbr.2018.02.017.

Folego G, Weiler M, Casseb RF, Pires R, Rocha A. Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI. Frontiers in Bioengineering and Biotechnology. 2020; 8:1193. https://www.frontiersin.org/article/10.3389/fbioe.2020.534592, doi: 10.3389/fbioe.2020.534592.

He K, Zhang X, Ren S, Sun J, Deep Residual Learning for Image Recognition; 2015.

Hwang H, Rehman HZU, Lee S. 3D U-Net for Skull Stripping in Brain MRI. Applied Sciences. 2019; 9(3). https://www.mdpi.com/2076-3417/9/3/569, doi: 10.3390/app9030569.

Jung W, Jun E, Suk HI. Deep recurrent model for individualized prediction of Alzheimer's disease progression. Neuroimage. 2021 08; 237:118143.

Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog): Modifications and Responsiveness in Pre-Dementia Populations. A Narrative Review. J Alzheimers Dis. 2018; 63(2):423–444.

Kumar S, Oh I, Schindler S, Lai AM, Payne PRO, Gupta A. Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. JAMIA Open. 2021 Jul; 4(3):ooab052. [PubMed Central:PMC8327375] [DOI:10.1093/jamiaopen/ooab052] [PubMed:29054264].

Lee B, Yamanakkanavar N, Choi JY. Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture. PLoS One. 2020; 15(8):e0236493.

Llano DA, Laforet G, Devanarayan V. Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to Alzheimer disease. Alzheimer Dis Assoc Disord. 2011; 25(1):73–84.

Martinez-Murcia FJ, Ortiz A, Gorriz JM, Ramirez J, Castillo-Barnes D. Studying the Manifold Structure of Alzheimer's Disease: A Deep Learning Approach Using Convolutional Autoencoders. IEEE J Biomed Health Inform. 2020 01; 24(1):17–26.

Mofrad SA, Lundervold AJ, Vik A, Lundervold AS. Cognitive and MRI trajectories for prediction of Alzheimer's disease. Sci Rep. 2021 Jan; 11(1):2122.

Nalepa J, Marcinkiewicz M, Kawulok M. Data Augmentation for Brain-Tumor Segmentation: A Review. Front Comput Neurosci. 2019; 13:83.

Popescu SG, Glocker B, Sharp DJ, Cole JH. Local brain-age: A U-Net model. bioRxiv. 2021; https://www.biorxiv.org/content/early/2021/08/04/2021.01.26.428243, doi: 10.1101/2021.01.26.428243.

Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry. 1984 Nov; 141(11):1356–1364.

Sabbagh MN, Malek-Ahmadi M, Kataria R, Belden CM, Connor DJ, Pearson C, Jacobson S, Davis K, Yaari R, Singh U. The Alzheimer's questionnaire: a proof of concept study for a new informant-based dementia assessment. J Alzheimers Dis. 2010; 22(3):1015–1021.

Stonnington CM, Chu C, Klöppel S, Jack CR, Ashburner J, Frackowiak RSJ. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage. 2010; 51(4):1405–1413. https://www.sciencedirect.com/science/article/pii/S1053811910003381, doi: https://doi.org/10.1016/j.neuroimage.2010.03.051.

Stout JC, Jernigan TL, Archibald SL, Salmon DP. Association of dementia severity with cortical gray matter and abnormal white matter volumes in dementia of the Alzheimer type. Arch Neurol. 1996 Aug; 53(8):742–749.

Zandifar A, Fonov VS, Ducharme S, Belleville S, Collins DL. MRI and cognitive scores complement each other to accurately predict Alzheimer's dementia 2 to 7 years before clinical onset. Neuroimage Clin. 2020; 25:102121.

Zhang J, Li Q, Caselli RJ, Ye J, Wang Y, Multi-task Dictionary Learning based Convolutional Neural Network for Computer aided Diagnosis with Longitudinal Images; 2017.

Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging. 2001 Jan; 20(1):45–57.