ELSEVIER

Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring ■ (2018) 1-13

Alzheimer's & Dementia

Diagnostic Assessment & Prognosis

Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment

Shangran Qiu[a], Gary H. Chang[b], Marcello Panagia[c,d], Deepa M. Gopal[c,d], Rhoda Au[e,f,g,h,i], Vijaya B. Kolachalama[b,d,j,*]

[a]Department of Physics, College of Arts and Sciences, Boston University, Boston, MA, USA
[b]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA
[c]Section of Cardiovascular Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA
[d]Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA, USA
[e]The Framingham Heart Study, Boston University School of Medicine, Boston, MA, USA
[f]Department of Anatomy and Neurobiology, Boston University School of Medicine, Boston, MA, USA
[g]Department of Neurology, Boston University School of Medicine, Boston, MA, USA
[h]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA
[i]Boston University Alzheimer's Disease Center and Boston University CTE Center, Boston University School of Medicine, Boston, MA, USA
[j]Hariri Institute for Computing and Computational Science and Engineering, Boston University, Boston, MA, USA

Abstract

Introduction: Our aim was to investigate if the accuracy of diagnosing mild cognitive impairment (MCI) using the Mini–Mental State Examination (MMSE) and logical memory (LM) test could be enhanced by adding MRI data.

Methods: Data of individuals with normal cognition and MCI were obtained from the National Alzheimer Coordinating Center database (n = 386). Deep learning models trained on MRI slices were combined to generate a fused MRI model using different voting techniques to predict normal cognition versus MCI. Two multilayer perceptron (MLP) models were developed with MMSE and LM test results. Finally, the fused MRI model and the MLP models were combined using majority voting.

Results: The fusion model was superior to the individual models alone and achieved an overall accuracy of 90.9%.

Discussion: This study is a proof of principle that multimodal fusion of models developed using MRI scans, MMSE, and LM test data is feasible and can better predict MCI.

Keywords: MRI; MMSE; LM test; Mild cognitive impairment; Deep learning; Convolutional neural network; Multilayer perceptron; Majority voting

1. Introduction

Cognitive decline is one of the most concerning behavioral symptoms associated with Alzheimer's disease (AD). Seamless changes in the AD continuum take years if not decades to progress from normal cognition (NC) to mild cognitive impairment (MCI), with gradual evolution of clinically probable AD to confirmed AD [1–3]. Early detection and accurate diagnosis of AD require careful medical assessment, including patient history as well as physical and neurological examinations. The Mini–Mental State Examination (MMSE), which is a 30-point questionnaire [4–6], is a brief cognitive assessment tool commonly used to screen for dementia, and the Wechsler Memory Scale Logical memory (LM) test is widely used to assess verbal memory and is considered a sensitive test of AD [7].

*Corresponding author. Tel.: +617-358-7253; Fax: +617-414-3292.
E-mail address: vkola@bu.edu

https://doi.org/10.1016/j.dadm.2018.08.013

Neuroimaging modalities such as magnetic resonance imaging (MRI) have shown to provide biologic evidence that cognitive decline is neurodegenerative [8–22], as they contain detailed information regarding the subcortical structures, good contrast of the gray matter, and the integrity of the brain tissue. Specifically, it is known that changes including cortical damage, focal lesions, and gray matter loss in the occipital, parietal, prefrontal, and temporal lobes can be understood using MRI scans [23]. Cortical atrophy is closely related to dementia and cognitive decline. For example, posterior cortical atrophy (PCA) is considered as a variant of AD in some studies [24–27], and this might spread to other brain regions that are commonly damaged in AD patients [28]. Brain MRI scans are characterized as complex, unstructured data structures and thus require sophisticated means by which to perform efficient, quantitative analysis. Although the diagnostic and psychometric strengths and limitations of neuropsychological tests and MRI scans have been tried and tested, there is a limited body of work that has attempted to understand the cumulative impact of combining these data sets for diagnosis of MCI.

Machine learning techniques have been widely used over the past few years for the analysis of biomedical imaging data and nonimaging data. More recently, a machine learning framework known as deep learning, which is based on artificial neural networks, has received increased attention because of its remarkable success in predicting various clinical outcomes of interest. In subspecialties that rely on imaging data, scientists have developed convolutional neural network (CNN) models, which are efficient deep learning techniques for object recognition and classification [29]. CNN models have proven ability to detect patterns within image data sets and predict corresponding outputs for common tasks such as classification with superior accuracy. Inside a CNN, a series of filters, with a size equivalent to a small image patch, automatically search through the whole image to find similar spatial features of the image. These filters can be learned and independently updated, so that a collection of them can detect critical information for a specific task and data set. Such approaches can be used in a relatively straightforward fashion to train deep learning models on 2D MRI scans. Other forms of deep neural networks based on a multilayer perceptron (MLP) architecture can be used to train nonimaging data sets such as the ones based on the MMSE and LM test results. Although these models can be informative on their own, a framework to combine all these models and the information contained therein will enhance their clinical utility. Fusion of similar forms of data has been previously performed to improve the prediction for the diagnosis of cognitive status. For example, different image projections, for example, axial, sagittal, and coronal, were combined to generate a unified model that was associated with MCI [30]. Structural MRI was also combined with positron emission tomography for similar purposes [31–33]. However, there is a limited body of work that has investigated how to efficiently combine MRI data with other nonimaging data sets such as the ones derived from the MMSE and LM tests.

The ability to efficiently distinguish individuals with MCI from the ones who have NC is crucial within the realm of early detection of AD, as the changes that need to be captured may only be subtle. Our hypothesis is that combining structural information derived from neuroimaging data and functional information derived from well-known screening tools and cognitive assessment methods can result in a better combined metric of diagnosing MCI.

2. Methods

Our goal was to create a predictive model of MCI by considering detailed structural and anatomic information contained within the MRI images as well as cognitive function assessed using the MMSE and LM tests. We first developed three individual machine learning models, one using MRI scans alone, another using only MMSE results, and third using only LM test results. Later, we combined the prediction of these models using majority voting, thus exemplifying multimodal data fusion.

2.1. Study participants and measures

Our proof-of-principle study was conducted using the data provided by the National Alzheimer's Coordinating Center (NACC), which was established by the National Institute on Aging/NIH in 1999 to facilitate collaborative AD research. Specifically, we selected a cross-sectional collection of individuals from the NACC data set (Table 1), with one MRI scan from each participant based on the following criteria: (1) participants were clinically diagnosed as either NC or MCI; (2) participants had both LM and MMSE test scores; (3) MRI scans with at least 20 slices were taken in the axial plane; (4) type of MRI was either T1 weighted or FLAIR as they both are fluid suppressed with dark cerebrospinal fluid

Table 1
Characteristics of the study population

Label	NC	MCI
Number of patients	303	83
Number of MRI scans	303	83
Age—median (range)	60 (43–89)	75 (56–87)
Percent male	34.3	61.4
Education—median (range)	16 (8–25)	16 (8–20)
LM test—median (range)	14 (3–22)	7 (0–16)
Total MMSE score—median (range)	30 (24–30)	27 (17–30)

Abbreviations: MRI, magnetic resonance imaging; LM, logical memory; MMSE, Mini–Mental Status Examination; NC, normal cognition; MCI, mild cognitive impairment.

NOTE. Data were obtained from the National Alzheimer's Coordinating Center (NACC) database. Both individuals with MCI (83) and the ones with NC (303) were considered. Both the MMSE and LM test results were also obtained and reported. Note that few individuals underwent MRI scanning more than once. Only one MRI scan was selected from each patient. The MMSE and LM test were conducted closer to the scan time.

and relative bright white and gray matter [15]; and (e) dimension of each cross-sectional slice of the MRI (height or width) was about 256 pixels so as to make this slice amenable for training the machine learning model. This search resulted in the selection of 386 unique cases. Among them, there were 303 individuals who were diagnosed as NC and 83 individuals who were diagnosed as MCI (82 amnestic MCI and one non-amnestic MCI).

Each of the 386 MRI scans had a variable number of 2D cross-sectional images (or slices), and the location and orientation of the brain within each scan were also different. To utilize each of these scans, we first grouped all 2D slices per participant in 20 bins and then selected the first 2D slice from each bin. We then manually selected a "signature" slice from the 20 slices for each scan merely based on highest similarity of anatomical features without knowing any information about clinical diagnosis that was going to be predicted. This "signature" slice was considered as slice 1 for modeling (Fig. 1) and largely covered cross-sections of the occipital horn, frontal horn of lateral ventricle, and thalamus. The next two adjacent slices in the superior direction were selected as slice 2 and slice 3, respectively. As a result, the selected slices covered various regions including lateral ventricles, inferior temporal, and middle temporal cortices. All these anatomic areas were previously reported as regions of interest that correlated with AD and MCI [34].

We selected three features including MMSEORDA (orientation subscale score—time), MMSEORLO (orientation subscale score—place), and NACCMMSE (total MMSE score) to generate a model based on MMSE results alone. A previous study has shown that MMSE orientation (e.g., time or place) was the domain with the largest extent of change over time in AD patients [35], and two other studies found that MMSE orientation is impaired early in the disease process [36,37]. Also, we selected all the features that are considered as part of the LM test to generate an LM test–based model. This includes LOGIPREV (total score from the previous test administration), LOGIMEM (total number of story units recalled from this current test administration), MEMUNITS (total number of story units recalled), and MEMTIME (time elapsed since first recall to delayed recall). The numeric value of the data was derived from the NACC Uniform Data Set Researcher's Data Dictionary (version 3.0, March 2015).

2.2. Training and validation of the MRI models

For each 2D MRI slice, we adapted the Oxford University's Visual Geometry Group's (VGG-11) model [38], pretrained on millions of images with 1000 object classes by incorporating minor changes to fine-tune the framework and to associate MRI image features with the clinical diagnosis of cognitive status (Fig. 2, Supplementary Table 1). Specifically, we inserted the batch normalization layer after every convolutional layer and a dropout layer after every max pooling layer within the VGG-11 architecture. We also added two fully-connected (FC) layers after the output layer of VGG-11 to perform binary classification on the output of VGG-11 that formed the deep neural network. After the first newly added FC layer, we added a dropout layer with rectified linear unit activation, and a softmax function was applied on the second FC layer. A similar approach

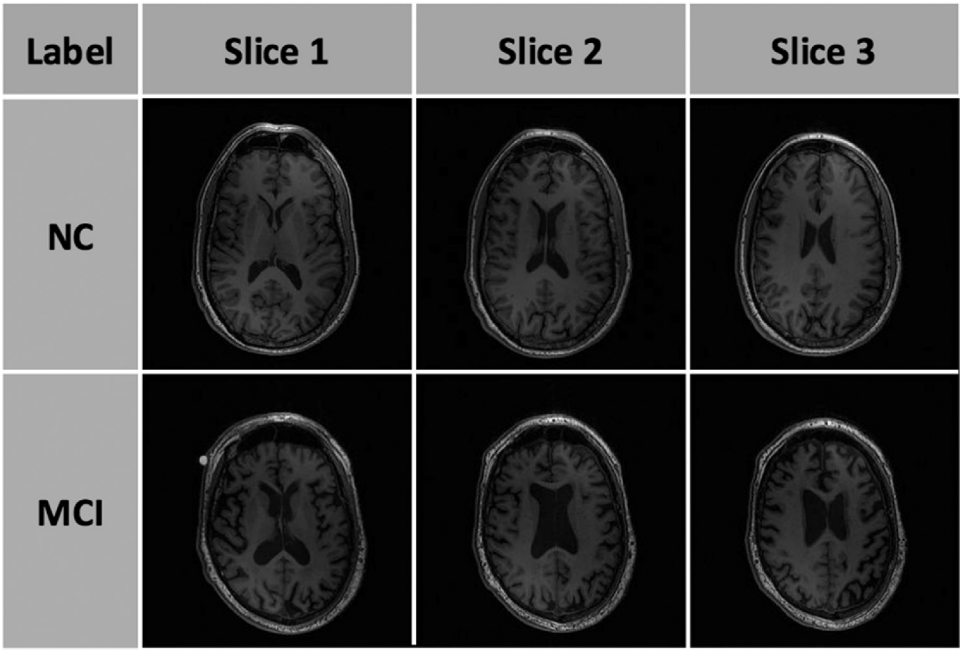

Fig. 1. Axial slices from different locations for two individuals with NC and MCI, respectively, are shown. The locations were carefully selected after processing the entire scans as they represented locations that have been shown to associate MCI [34]. Abbreviations: MCI, mild cognitive impairment; NC, normal cognition.

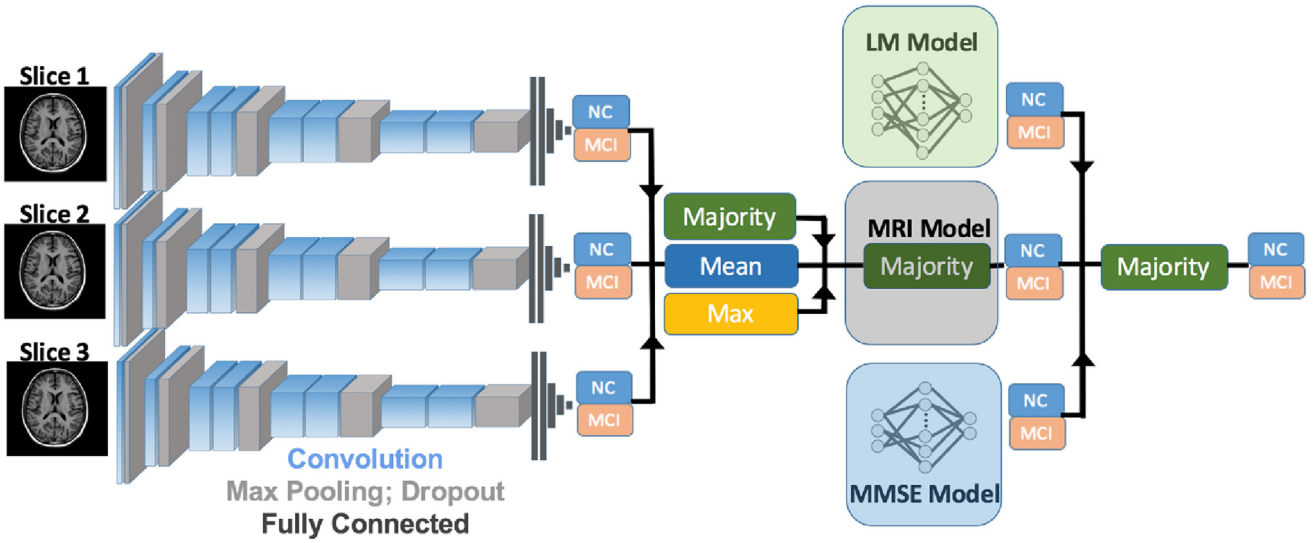

Fig. 2. Schematic of the modeling architecture. Three deep neural network models (VGG-11) were independently trained and tested on three slices, respectively. Max, mean, and majority voting were performed on the outputs of the three VGG-11 models. Predictions from the three voting methods were then combined by majority voting again to generate a fused MRI model. Two MLP models were independently developed on MMSE and LM test results, respectively. Finally, predictions from the three base models (MRI model, MMSE model, and LM model) were combined by applying another majority voting to generate a final prediction of multimodal fusion model. Abbreviations: MLP, multilayer perceptron; MCI, mild cognitive impairment; NC, normal cognition; MRI, magnetic resonance imaging; LM, logical memory; MMSE, Mini–Mental Status Examination.

was developed on another well-known CNN architecture (VGG-16), where dropout layers were added to avoid model overfitting [39].

The original data were split in 2:1:1 for training, validation, and testing. Care was taken to ensure that each split had the same proportion of NC and MCI cases. The original MRI slices were cropped to 224 × 224 pixels and were normalized to 0 to 1 to make them compatible with the input requirements of the pretrained VGG-11 model. For completeness, we performed model training and testing 5 different times. Mean and standard deviation were calculated for each performance metric over the five iterations to demonstrate model behavior due to random splitting. We set the class-specific weighted cross-entropy as the loss function to be optimized through the stochastic gradient descent algorithm [40]. The explicit form of the loss function is as follows:

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\left[-w_{MCI}\,\widehat{y}_i \log(y_i) - w_{NC}\left(1-\widehat{y}_i\right)\log(1-y_i)\right], \tag{1}$$

where $\widehat{y}_i$ is the label of the i^{th} scan, which takes value 0 if the i^{th} scan is associated with NC and 1 if the i^{th} scan is associated with MCI. y_i and $1-y_i$ are the probabilities, predicted by the model, that the i^{th} scan is associated with MCI or NC, respectively. The class-specific weights, w_{MCI} and w_{NC}, can be tuned as hyperparameters to alleviate the influence of data imbalance. The value of $w_{MCI}$and w_{NC} was set at 4 and 1, respectively, in our MRI models.

We performed model training using a transfer learning strategy, which is a well-known approach to train image-based classification problems [41,42]. We froze all the pretrained VGG-11 parameters and trained the externally added FC layers on the training data. During the training process, the parameters of the FC layers were saved when the F1 score calculated on the validation set was at maximum. Then, we performed fine-tuning of the deep neural network after the saved parameters of the two FC layers were loaded into the model. The network with highest F1 score on validation set was saved and used for testing. For the aforementioned two stages of training (training the FC layers and full network fine-tuning), we used a batch size of 20 using backpropagation with a gradient descent optimizer whose momentum was set to 0.9 and the learning rate was set to 0.05.

Both dropout and batch normalization layers can efficiently decrease the risk of overfitting and increase the model generalizability. During training, nodes in front of dropout layers were randomly deactivated by dropout according to the probability "p" [43]. Thus, only part of the model parameters selected randomly were updated by the gradient descent algorithm. With batch normalization, output from the previous layers of batch normalization was scaled and shifted by

$$y_i = \gamma\,\frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta, \tag{2}$$

where x_i is the i^{th} element of the output, and μ_B and σ_B^2 are the mean and variance of x_i calculated over a batch [44]. The scale parameter γ and shift parameter β were learned through the training process along with other parameters within the model to reduce internal covariate shift and stabilize training process.

2.3. Training and validation of the MMSE and LM models

MMSE and LM cutoffs used in current AD clinical trials and diagnostic studies have limited diagnostic accuracy, particularly for distinguishing between normal cognition and MCI, and MCI from AD dementia [45]. The total MMSE score or delayed LM score does not contain all the available information within the MMSE and LM test results. Thus, we developed two independent MLP models taking into account the subscores within the MMSE and LM test, respectively, to predict MCI.

Because the input layer can have any number of nodes, we can take all available features from the MMSE and LM test to build MLP models for MMSE and LM. The combination of linear transformation and nonlinear activation functions makes MLP a good choice that is capable of exploring complex underlying relationships between selected examination features and the diagnosis of MCI.

Both MMSE and LM models have one input layer in which the MMSE model has 3 nodes and LM model has 4 nodes, one hidden layer with 20 nodes and one output layer with two nodes. In addition, sigmoid and softmax nonlinear functions were added at the hidden layer and output layer, respectively. These models were trained using backpropagation with a gradient descent optimizer whose momentum was set to 0.9 with learning rate 0.05. Loss function used for training the models was the same as defined in Equation 2. The class-specific weights, w_{MCI} and w_{NC}, were set at 4 and 1 for LM model and was set at 5 and 1 for MMSE model, respectively. The model with the highest F1 score on validation set was saved during the training process and used to further evaluate performance on the testing set.

2.4. Model fusion and visualization of majority voting

The predictions from these trained VGG models were then combined using a series of voting approaches. Specifically, we performed majority voting, max voting, and mean voting on three independent modified VGG-11 models (Fig. 2). Majority voting was performed again on the predictions of these voting approaches on the test data to finally generate an image-based fusion model. Mean voting takes the mean probabilities predicted from each modified VGG-11 model trained on the three slices as the final prediction defined as

$$P_{NC}^{mean} = \left(P_{NC}^{1} + P_{NC}^{2} + P_{NC}^{3}\right)/3 \tag{3}$$

$$P_{MCI}^{mean} = \left(P_{MCI}^{1} + P_{MCI}^{2} + P_{MCI}^{3}\right)/3$$

Here, P_{NC}^{i} and P_{MCI}^{i} are the probabilities predicted from the i^{th} model that one specific scan belongs to NC or MCI, respectively. Max voting takes the prediction from the model with the largest confidence as final prediction, where confidence is defined as $max(P_{NC}^{i}, P_{MCI}^{i})$, because the model can predict either NC or MCI with high confidence. Majority voting takes the label, NC or MCI, which was predicted for equal or more than 2 times as the final prediction. Then, using the fused MRI model based on the three independent 2D scans along with the MLP models generated with MMSE and LM test results, respectively, we performed majority voting to generate a multimodal fusion model to predict MCI.

To visualize and evaluate the performance of the multimodal fusion model, we also performed a subgroup analysis where cases were split into eight categories defined based on the predictions made by the fused MRI, MMSE, and LM model, respectively. Within each subgroup, the accuracy of the final prediction made from applying majority voting on the fused MRI model, the MMSE model, and the LM model were also visualized into eight clusters.

2.5. Model runs and performance metrics

Considering the small size of the data set, model runs for each case (fused MRI model, MMSE model, LM model, and fused multimodal model) were performed 5 times where the data were split randomly into training, validation, and testing for each run. Model performance was evaluated by averaging various performance metrics across the 5 runs. Accuracy, precision, recall, F1 score, area under the curve (AUC or c-statistic), and Matthews correlation coefficient (MCC) were reported as mean values along with standard deviation.

3. Results

Data from the NACC database allowed us to develop a fused multimodal machine learning model that had the ability to accurately perform diagnosis of MCI. This task required extensive amount of preprocessing that involved manually observing the appropriate cross-sectional MRI slices that were selected for developing the model.

3.1. MRI model

The performances of the three independent VGG-11 models that were trained and evaluated on 3 slices, respectively, are shown in Table 2. The receiver operating characteristic curves for each model, with mean AUC 0.833, 0.827, and 0.844, respectively, are shown in Fig. 3. Similar model performances on the three slices indicate that the VGG-11 architecture was able to capture the nuances of the 2D image regions in a consistent fashion. Moreover, each image represented a unique location within the brain that presumably was able to correlate with the corresponding diagnosis with high accuracy. A series of voting techniques were performed to generate the consistent prediction from the MRI scans. We can see that the fused MRI model has similar performance as the model trained on slice 1 and outperformed the rest of the two models that were trained on slices 2 and 3 (Fig. 4, Table 2 & Table 3).

Table 2
Performance metrics for the three modified VGG-11 models

Slice number	Accuracy	Precision	Recall	F1	Matthews correlation coefficient
Slice 1	83.3 ± 4.1%	0.876 ± 0.036	0.918 ± 0.037	0.896 ± 0.025	0.476 ± 0.142
Slice 2	80.8 ± 3.6%	0.844 ± 0.023	0.929 ± 0.059	0.883 ± 0.025	0.378 ± 0.116
Slice 3	82.1 ± 2.2%	0.893 ± 0.027	0.879 ± 0.052	0.884 ± 0.017	0.490 ± 0.056

NOTE. We developed three modified VGG-11 models on three slices, respectively. Metrics are shown on the testing data set (n = 97) that was not used for model training.

3.2. MMSE and LM models

The performance metrics of the MMSE model and LM model, constructed on the training data set and evaluated on the testing data sets, are shown in Table 3. The receiver operating characteristic curves for the MMSE model and the LM model are shown in Fig. 5. We observed that the LM model has better performance than the MMSE model and MRI model in terms of accuracy, precision, F1, and MCC. However, the MMSE and MRI models have their own strengths, such as higher recall value than that of the LM model. These differences highlight the power of each test, and the models are developed using individual tests alone. They also point to the importance of fusing these models so as to be able to gain additional insight on cognitive function.

3.3. Multimodal fusion model

Given the predictions from the three base models, that is, the MRI model, the MMSE model, and the LM model, final prediction of the fusion model was performed using majority voting on the predictions of three base models (Fig. 6, Table 3). Specifically, the accuracy of the multimodal fusion model outperformed each base model by 7.8%, 6.6%, and 1.8%, respectively. The predictive performance of the fused multimodal model outperformed that of the three base models in accuracy, recall, F1 score, and MCC. The precision of the fused multimodal model is slightly lower than that of the LM model.

We also performed a subgroup analysis to further evaluate the performance of the multimodal fusion model (Fig. 7). All the cases in the test data were divided into eight clusters, depending on predictions made by each base model. To avoid any overlap, each case was assigned a random location around the related cluster. To visualize the accuracy of majority voting within each subgroup, the dot was presented as a blue circle if majority voting on predictions from three base models agreed with the true label, otherwise the dot was represented as a red triangle. For example, cluster E denotes all cases where both the MRI and the LM models predicted the outputs as NC, but the MMSE model predicted the output to be MCI. But after majority voting, the multimodal fusion model predicted the NC label because 2 of 3 base models voted NC. Among the 21 cases that belonged to this category, majority voting made 18 right decisions by correcting the mistakes made by the MMSE model and 3 wrong decisions as both the MRI model and the LM model incorrectly predicted that individual to have NC. Also, for the case of cluster G, when the MMSE and LM models predicted the individual to have MCI, the MRI model, however, predicted the individual to have NC. When majority voting was performed, the multimodal fusion model predicted the MCI label because 2 of 3 base models voted MCI. Among the 29 cases that belonged to this category, majority voting made 24 correct decisions by correcting the mistakes made by the MRI model and 5 wrong decisions as both MMSE model and the LM model incorrectly predicted that individual to have MCI. The accuracy of majority voting in each cluster was computed as (A) 97.7%, (B) 82.8%, (C) 62.5%, (D) 90.5%, (E) 85.7%, (F) 42.9%, (G) 82.8%, and (H) 90.6%. Taken together, these results indicate that majority voting on the three base models enhanced the accuracy of the multimodal fusion model in terms of predicting MCI.

In summary, from the three base models developed on MRI scans, MMSE, and LM test results, respectively, we observed that the LM model outperformed the fused MRI model and MMSE model, as revealed by most performance metrics. However, the MMSE and MRI models both have higher recall value than the LM model based on our experiments. We then performed the model fusion using majority voting on the three base models, as this strategy carefully generated the best performing model whose predictive performance had higher accuracy, recall, F1 score, and MCC than what any of the single models could achieve by themselves (Fig. 7).

4. Discussion

We explored the effectiveness of combining various forms of data with the underlying assumption that additional information can be gained from multiple data modalities to better assess MCI. Although previous studies have focused on combining different sources of imaging data such as fusing MRI, fMRI, and positron emission tomography scans, there is little body of work that focused on combining imaging data with other sources of information that can be obtained through well-established screening tests and neuropsychological examinations (i.e., MMSE and LM test). Our multimodal fusion model results demonstrate that multiple neural network models can be trained and combined together to generate a fused model to predict MCI.

The National Institute on Aging and the Alzheimer's Association built a workgroup with the task of revising the

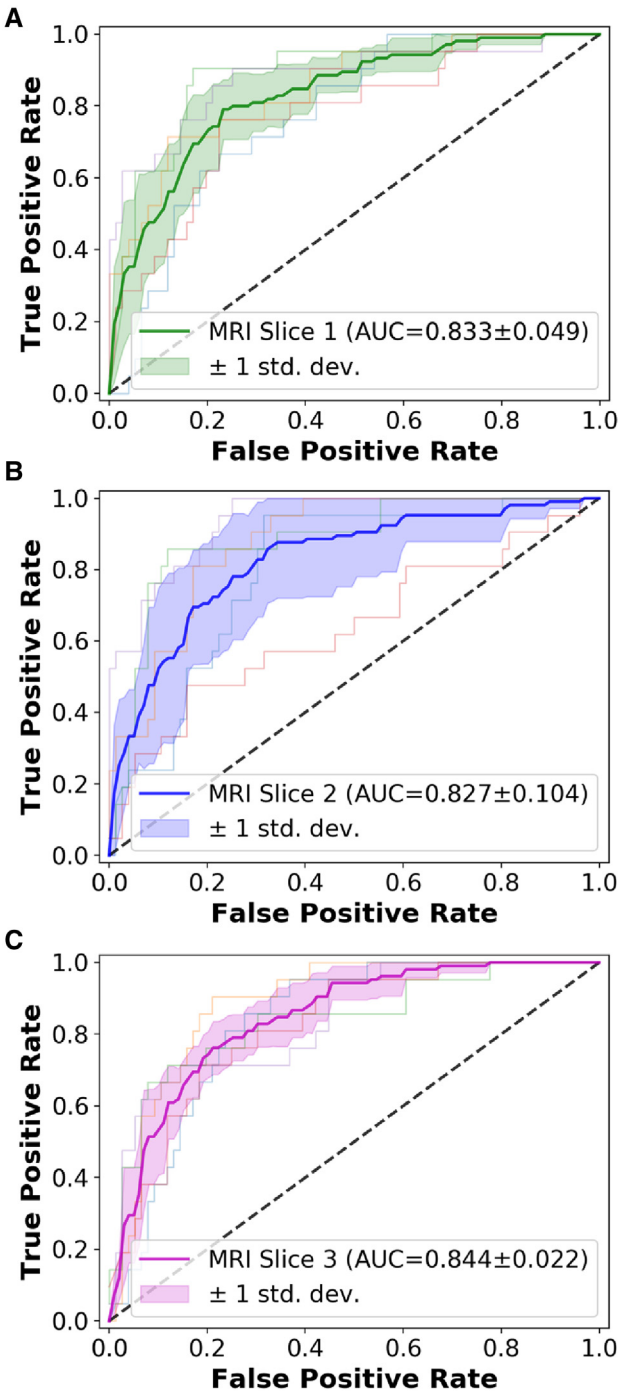

Fig. 3. ROC curves of three modified VGG-11 models. ROC curves of the modified VGG-11 model developed on the first (A), second (B), and third (C) slices, respectively. Each model was trained and evaluated on five different data splits. ROC curves for the five random splits are shown as five thin and transparent lines with different colors. The mean ROC curve and the standard deviation were shown as the bold line and shaded region, respectively. Abbreviations: AUC, area under the curve; MRI, magnetic resonance imaging.

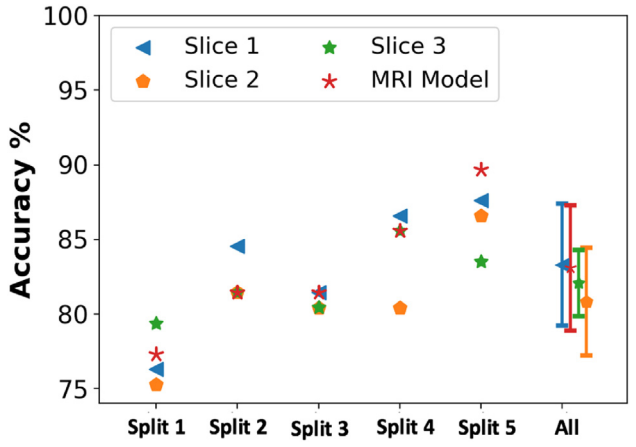

Fig. 4. Fusion of three modified VGG-11 models. Accuracies of each 2D MRI models developed on single slice and MRI-fused model from a series of voting methods for five random splits are shown. Mean and standard deviation over five splits for each model's accuracy are also shown, respectively.

1984 criteria for AD dementia and MCI due to AD [46,47]. The core clinical criteria for AD and MCI will continue to be the cornerstone of the clinical diagnosis, but biomarker evidence is expected to improve the pathophysiological accuracy of the diagnosis in the research setting. One implication in the original criteria of AD dementia is that memory impairment is always the primary cognitive impairment in all patients. However, it is critical to examine cognitive domains in addition to memory that might be impaired. These cognitive domains include language, visuospatial skills, executive functions, and attentional control. Our motivation to combine the predictions from the MMSE model with the LM model is to comprehensively consider the cognitive impairment in various domains. Biomarkers can be used to assist and enhance the clinical diagnosis of AD and MCI. The progress of the gradual cognitive decline can happen along with a wide range of structural changes in the brain, including frontotemporal lobar degeneration associated with language disorder and posterior cortical atrophy relevant to visual disability. However, considering MCI is not biomarker confirmed, we designed the deep learning model to explore the differences between participants with NC and MCI automatically from the data set. We expected the MRI model, carrying comprehensive structural information of the brain, can further push the diagnostic accuracy from MMSE and LM models to a superior level.

4.1. Data sets for model development

Compared with other brain MRI data sets, including Alzheimer's Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS), the NACC has its own strengths and limitations. The LM test provided well-validated verbal memory data that were sufficient for these proof-of-concept analyses. ADNI and OASIS use list learning tests for their measure of verbal memory, but the data present greater complexity for this type of analysis. Thus, the NACC data set was chosen to achieve the goals

Table 3
Performance metrics for the 3 base models along with the multimodal fusion model

Model type	Accuracy	Precision	Recall	F1	MCC
MRI model	83.1 ± 4.2%	0.878 ± 0.031	0.913 ± 0.050	0.894 ± 0.027	0.481 ± 0.132
MMSE model	84.3 ± 2.3%	0.888 ± 0.044	0.921 ± 0.049	0.902 ± 0.014	0.518 ± 0.111
LM model	89.1 ± 1.9%	**0.951 ± 0.024**	0.908 ± 0.001	0.929 ± 0.011	0.698 ± 0.067
Majority voting	**90.9 ± 2.7%**	0.926 ± 0.037	**0.963 ± 0.015**	**0.944 ± 0.015**	**0.719 ± 0.101**

Abbreviations: MRI, magnetic resonance imaging; LM, logical memory; MMSE, Mini–Mental Status Examination; MCC, Matthews correlation coefficient.
NOTE. Three base models were developed independently on MRI slices, MMSE features, and LM features, respectively. The multimodal fusion model combined three base models using majority voting. Model performance was evaluated on testing set (n = 97) that was not used in training.
NOTE. Bold font indicates the best value in that column.

of this study. MRI scans included in ADNI and OASIS data sets were collected with consistent MRI settings, so that the size and contrast over different MRI scans are consistent. However, MRI scans in NACC data sets were collected from different NIA-funded Alzheimer's Disease Centers across the United States. The inconsistency of the MRI settings in different centers made the procedure of analyzing MRI scans more complicated. First, owing to the fact that the dimensions of MRI are all different, we could not apply automatic 3D brain registration techniques on those MRI scans whose number of slices ranges from 20 to more than 200. In addition, the inconsistency made developing 3D CNN on the NACC MRI data extremely challenging because the CNN model used in classification task needs to take fixed-sized data as input. To diagnose those with only limited slices within an MRI scan, fusion of 2D CNN models can be an interesting choice. Thus, we manually selected 3 slices from each scan based on anatomical landmarks to decrease the effect of misalignment, that is, inconsistent location and orientation of brains within each scan. Future studies are warranted to test our approach on other cohort studies, such as ADNI and OASIS.

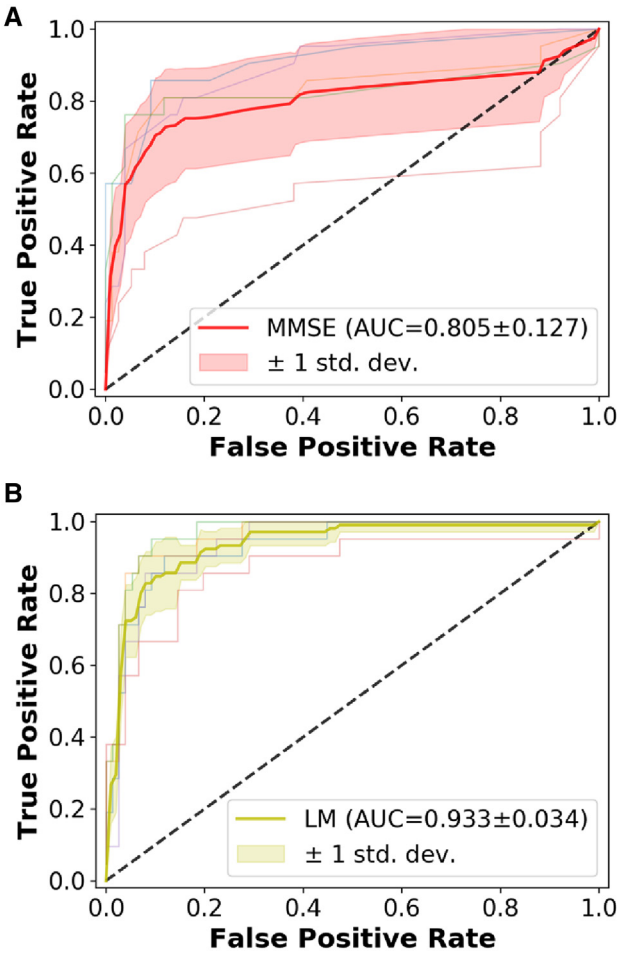

Fig. 5. ROC curves for the MMSE (A) and LM test (B) models. Each model was trained and evaluated on five different data splits. ROC curves for 5 random splits are shown as 5 thin and transparent lines with different colors. The mean ROC curve and the standard deviation are shown as the bold line and the shaded region, respectively. Abbreviations: AUC, area under the curve; LM, logical memory; MMSE, Mini–Mental Status Examination.

4.2. Role of deep learning

Biomedical image analysis is one of the successful beneficiaries of the data science revolution over the past decade [48]. Sophisticated hardware along with cutting edge machine learning algorithms including the deep learning frameworks

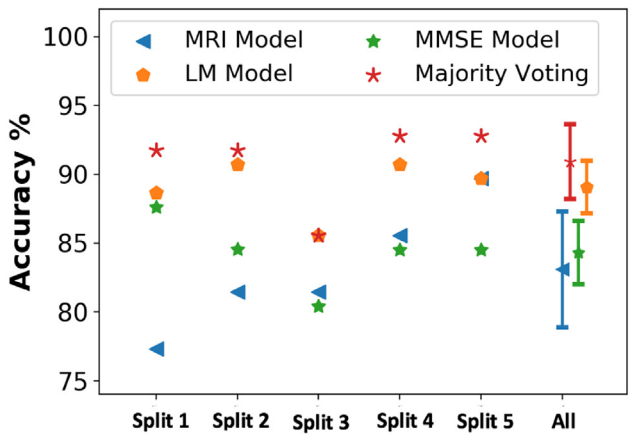

Fig. 6. Performance of majority voting on the 3 base models. Accuracy of each base model (MRI model, MMSE model, and LM model) and multimodal fusion model for five random splits are shown. Mean and standard deviation over five splits for each model's accuracy are also shown. Abbreviations: MRI, magnetic resonance imaging; LM, logical memory; MMSE, Mini–Mental Status Examination.

web 4C/FPO

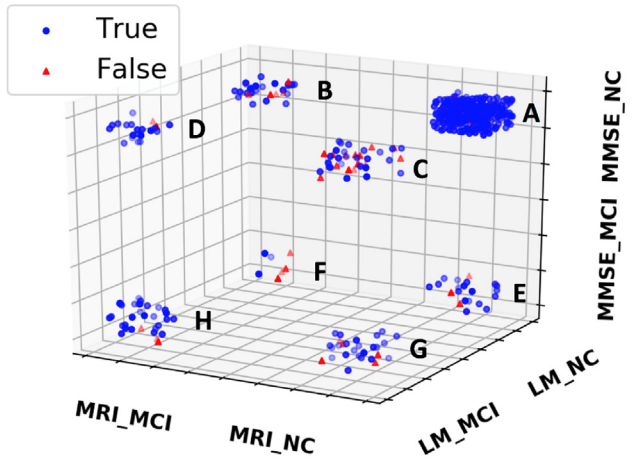

Fig. 7. Subgroup analysis of the multimodal fusion model. (A) Cases predicted by three base models as NC. (B) Cases predicted by the MMSE model and LM model as NC but predicted as MCI by the MRI model. (C) Cases predicted by the MRI model and MMSE model as NC but predicted as MCI by the LM model. (D) Cases predicted by the MMSE model as NC but predicted as MCI by the MRI model and LM model. (E) Cases predicted by the MRI model and LM model as NC but predicted as MCI by the MMSE model. (F) Cases predicted by the LM model as NC but predicted as MCI by the MRI model and MMSE model. (G) Cases predicted by MRI model as NC but predicted as MCI by the MMSE model and LM model. (H) Cases predicted by three base models as MCI. Note that "True" denoted by a blue circle indicates that majority voting prediction matched with the true label on that case, and "False" denoted by a red triangle indicates that majority voting prediction did not match with the underlying label of that case. Abbreviations: MRI, magnetic resonance imaging; LM, logical memory; MMSE, Mini–Mental Status Examination; MCI, mild cognitive impairment; NC, normal cognition.

have allowed us to make remarkable progress in terms of developing diagnostic models for disease assessment and even predictive models for disease prognosis. Many researchers indeed believe that we are just at the beginning stages of this revolution as more advances in data-learning technologies can profoundly impact the way we make use of biomedical data. Given the right infrastructure, deep learning algorithms are now arguably the primary choice for image-based classification of clinical phenotypes as they have shown to be more effective than using traditional machine learning models for these tasks [49,50]. Significant progress has been made in the field of AD where both traditional machine learning and deep learning methodologies have been implemented to identify various patterns and perform segmentation within the brain as well as develop diagnostic models for accurate detection of clinical phenotypes.

The traditional way of training CNN models requires large amounts of imaging data as there are thousands to millions of parameters that need to be estimated to obtain reasonably accurate predictions using them. This approach is not practically feasible to analyze neuroimaging data because only a limited number of labeled cases are generally available. Moreover, training a CNN model *de novo* with limited cases can easily lead to model overfitting. Therefore, we leveraged transfer learning that allowed us to fine-tune the weights of a pretrained deep neural network, thus making it practically viable to use deep learning for data sets with limited cases [41,42].

We observed that adding dropout and batch normalization layers between convolutional blocks efficiently stabilized and accelerated the training process. Because we were dealing with a slightly imbalanced data set (NC-MCI ratio: 3.65), the model can get stuck in local minima during training. Increasing the learning rate is typically one trick to avoid this issue, but naively increasing the learning rate without batch normalization can cause gradients to explode or even vanish [44]. Thus, we performed batch normalization, rescaled, and shifted the outputs between the intermediate layers within the deep neural network, to allow the model to get trained with large learning rate in a stable fashion. We also added dropout layers to overcome overfitting, as this technique is well known to improve the generalizability of the model [43].

4.3. Study limitations and future directions

Incorporating biomarkers into the diagnosis of MCI has some limitations. There are a broad set of biomarkers correlated with AD including β-amyloid, cerebrospinal fluid–tau, fluorodeoxyglucose uptake, and cortical atrophy [47]. It is still not clear how to select a subset of biomarkers for specific diagnosis and how the information contained in these biomarkers should be collected and analyzed in a standardized fashion. Another limitation of this study is that we do not know whether the subjects diagnosed as MCI can be placed on the AD continuum because there is no confirmed biomarker for MCI due to AD. Moreover, it is known that Lewy body and cerebrovascular disease also cause impairment in cognition [47].

In future, one could envision the use of MRI in conjunction with clinical core assessments to provide more specific and accurate diagnosis of AD and other non-AD-related dementias. Neurobiological changes in the brain occur years before symptoms of impaired cognition appear, and patients with similar clinical symptoms of MCI may have substantial heterogeneity in biomarkers [51]. Thus, it is important to detect early stages of cognitive impairment and predict the future direction of the disease toward various dementias. Ultimately, we envision development of diagnostic software that can be built based on the deep learning models to automatically analyze various types of data. Models based on each data modality may result in specific predictions with bounded accuracies, and approaches such as majority voting can help unify these findings and even enhance the overall diagnostic accuracy.

5. Conclusion

By adding seemingly disparate information, such as MRI scans, to screening tools such as MMSE and neuropsychological examinations such as LM test, which are well

known and widely used to detect MCI in aging individuals, performance of deep-learning models for MCI diagnosis can be greatly enhanced. This result also suggests that no data modality may be sufficient to assess MCI on its own, and that this unique feature is even more salient when the goal is to diagnose diseases that have insidious onset such as Alzheimer's disease. For these models to reach mainstream clinical practice for real-time assessment of MCI, they need to be evaluated on multiple cohorts in the form of retrospective data analysis as well as prospective trials. If this happens, then the implications for clinical care are profound as they will likely lead to improved outcomes compared with the current methods for early detection of MCI.

Acknowledgments

This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through BU-CTSI Grant (1UL1TR001430), a Scientist Development Grant (17SDG33670323) from the American Heart Association, and a Hariri Research Award from the Hariri Institute for Computing and Computational Science & Engineering at Boston University to V.B.K., and NIH grants (R01-AG016495, R01-AG08122, R01-AG033040) to R.A. Additional support was provided by Boston University's Affinity Research Collaboratives program and Boston University Alzheimer's Disease Center (P30-AG013846). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P50 AG005134 (PI Bradley Hyman, MD, PhD), P50 AG016574 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI M. Marsel Mesulam, MD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P50 AG005131 (PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P50 AG033514 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), and P50 AG047270 (PI Stephen Strittmatter, MD, PhD). The authors also thank Ms. Nicole Thomas at the National Alzheimer's Coordinating Center, University of Washington, for her assistance with data transfer.

RESEARCH IN CONTEXT

1. Systematic review: After performing a PubMed search, we found that there is a limited body of work that investigated if the accuracy of diagnosing mild cognitive impairment (MCI) using well-known screening tools such as Mini–Mental State Examination and neuropsychological tests such as logical memory (LM) test could be enhanced by adding structural information from magnetic resonance imaging scans.
2. Interpretation: Our findings indicate that a fusion modeling framework can better predict MCI as it has the capability to combine needed information from multimodal data resources. This also implies that no data modality may be sufficient to assess MCI on its own.
3. Future directions: This framework when validated on multiple clinical cohorts can profoundly impact the way we make use of multimodal data sets for diagnosing individuals with MCI from the ones with normal cognition.

References

[1] Kelley BJ, Petersen RC. Alzheimer's disease and mild cognitive impairment. Neurol Clin 2007;25:577–609, v.
[2] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, et al. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:280–92.
[3] Aisen PS, Cummings J, Jack CR Jr, Morris JC, Sperling R, Frolich L, et al. On the path to 2025: understanding the Alzheimer's disease continuum. Alzheimers Res Ther 2017;9:60.
[4] Arevalo-Rodriguez I, Smailagic N, Roque IFM, Ciapponi A, Sanchez-Perez E, Giannakou A, et al. Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI). Cochrane Database Syst Rev 2015:CD010783.
[5] Harrell LE, Marson D, Chatterjee A, Parrish JA. The Severe Mini-Mental State Examination: a new neuropsychologic instrument for the bedside assessment of severely impaired patients with Alzheimer disease. Alzheimer Dis Assoc Disord 2000;14:168–75.
[6] Pangman VC, Sloan J, Guse L. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. Appl Nurs Res 2000;13:209–13.
[7] Wolf S. A compendium of neuropsychological tests: Administration, norms, and commentary. Integr Physiol Behav Sci 2000;35:70–1.

[8] Dickerson BC, Wolk DA, I. Alzheimer's Disease Neuroimaging. MRI cortical thickness biomarker predicts AD-like CSF and cognitive decline in normal adults. Neurology 2012;78:84–90.

[9] Godin O, Tzourio C, Rouaud O, Zhu Y, Maillard P, Pasquier F, et al. Joint effect of white matter lesions and hippocampal volumes on severity of cognitive decline: the 3C-Dijon MRI study. J Alzheimers Dis 2010;20:453–63.

[10] Jokinen H, Goncalves N, Vigario R, Lipsanen J, Fazekas F, Schmidt R, et al. Early-Stage White Matter Lesions Detected by Multispectral MRI Segmentation Predict Progressive Cognitive Decline. Front Neurosci 2015;9:455.

[11] Liem MK, Lesnik Oberstein SA, Haan J, Van Der Neut IL, Ferrari MD, Van Buchem MA, et al. MRI correlates of cognitive decline in CADASIL: a 7-year follow-up study. Neurology 2009;72:143–8.

[12] Mungas D, Harvey D, Reed BR, Jagust WJ, Decarli C, Beckett L, et al. Longitudinal volumetric MRI change and rate of cognitive decline. Neurology 2005;65:565–71.

[13] Mungas D, Reed BR, Jagust WJ, Decarli C, Mack WJ, Kramer JH, et al. Volumetric MRI predicts rate of cognitive decline related to AD and cerebrovascular disease. Neurology 2002;59:867–73.

[14] Reijmer YD, Van Den Berg E, De Bresser J, Kessels RP, Kappelle LJ, Algra A, et al. Accelerated cognitive decline in patients with type 2 diabetes: MRI correlates and risk factors. Diabetes Metab Res Rev 2011;27:195–202.

[15] Siraly E, Szabo A, Szita B, Kovacs V, Fodor Z, Marosi C, et al. Monitoring the early signs of cognitive decline in elderly by computer games: an MRI study. PLoS One 2015;10:e0117918.

[16] Swann A, O'brien J, Ames D, Schweitzer I, Desmond P, Tress B. Does hippocampal atrophy on MRI predict cognitive decline? A prospective follow-up study. Int J Geriatr Psychiatry 1997;12:1182–8.

[17] Tupler LA, Krishnan KR, Greenberg DL, Marcovina SM, Payne ME, Macfall JR, et al. Predicting memory decline in normal elderly: genetics, MRI, and cognitive reserve. Neurobiol Aging 2007;28:1644–56.

[18] Uiterwijk R, Van Oostenbrugge RJ, Huijts M, De Leeuw PW, Kroon AA, Staals J. Total Cerebral Small Vessel Disease MRI Score Is Associated with Cognitive Decline in Executive Function in Patients with Hypertension. Front Aging Neurosci 2016;8:301.

[19] Van Der Flier WM, Middelkoop HA, Weverling-Rijnsburger AW, Admiraal-Behloul F, Bollen EL, Westendorp RG, et al. Neuropsychological correlates of MRI measures in the continuum of cognitive decline at old age. Dement Geriatr Cogn Disord 2005;20:82–8.

[20] van der Flier WM, van Buchem MA, van Buchem HA. Volumetric MRI predicts rate of cognitive decline related to AD and cerebrovascular disease. Neurology 2003;60:1558. author reply 1558-9.

[21] Wright CB, Dong C, Caunca MR, Derosa J, Kuen Cheng Y, Rundek T, et al. MRI Markers Predict Cognitive Decline Assessed by Telephone Interview: The Northern Manhattan Study. Alzheimer Dis Assoc Disord 2017;31:34–40.

[22] Zhu M, Wang X, Gao W, Shi C, Ge H, Shen H, et al. Corpus callosum atrophy and cognitive decline in early Alzheimer's disease: longitudinal MRI study. Dement Geriatr Cogn Disord 2014;37:214–22.

[23] Lorenzetti V, Allen NB, Fornito A, Yucel M. Structural brain abnormalities in major depressive disorder: a selective review of recent MRI studies. J Affect Disord 2009;117:1–17.

[24] Ross SJ, Graham N, Stuart-Green L, Prins M, Xuereb J, Patterson K, et al. Progressive biparietal atrophy: an atypical presentation of Alzheimer's disease. J Neurol Neurosurg Psychiatry 1996;61:388–95.

[25] Aharon-Peretz J, Israel O, Goldsher D, Peretz A. Posterior cortical atrophy variants of Alzheimer's disease. Dement Geriatr Cogn Disord 1999;10:483–7.

[26] Kaskie B, Storandt M. Visuospatial deficit in dementia of the Alzheimer type. Arch Neurol 1995;52:422–5.

[27] Levine DN, Lee JM, Fisher CM. The visual variant of Alzheimer's disease: a clinicopathologic case study. Neurology 1993;43:305–13.

[28] Goethals M, Santens P. Posterior cortical atrophy. Two case reports and a review of the literature. Clin Neurol Neurosurg 2001;103:115–9.

[29] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

[30] Aderghal K, Benois-Pineau J, Afdel K. Classification of sMRI for Alzheimer's disease Diagnosis with CNN: Single Siamese Networks with 2D+epsilon Approach and Fusion on ADNI. In: ACM International Conference on Multimedia Retrieval; 2017. Bucharest, Romania: ACM; 2017.

[31] Vu TD, Yang HJ, Nguyen VQ, Oh AR, Kim MS. Multimodal learning using Convolution Neural Network and Sparse Autoencoder. 2017 Ieee International Conference on Big Data and Smart Computing. Bigcomp; 2017. p. 309–12.

[32] Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Health Inform 2018;22:173–83.

[33] Lei B, Chen S, Ni D, Wang T. Discriminative learning for Alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. Front Aging Neurosci 2016;8:77.

[34] Holland D, Brewer JB, Hagler DJ, Fennema-Notestine C, Dale AM, Alzheimer's Disease Neuroimaging I. Subregional neuroanatomical change as a biomarker for Alzheimer's disease. Proc Natl Acad Sci U S A 2009;106:20954–9.

[35] Small BJ, Viitanen M, Backman L. Mini-Mental State Examination item scores as predictors of Alzheimer's disease: incidence data from the Kungsholmen Project, Stockholm. J Gerontol A Biol Sci Med Sci 1997;52:M299–304.

[36] Galasko D, Klauber MR, Hofstetter CR, Salmon DP, Lasker B, Thal LJ. The Mini-Mental State Examination in the early diagnosis of Alzheimer's disease. Arch Neurol 1990;47:49–52.

[37] Fillenbaum GG, Wilkinson WE, Welsh KA, Mohs RC. Discrimination between stages of Alzheimer's disease with subsets of Mini-Mental State Examination items. An analysis of Consortium to Establish a Registry for Alzheimer's Disease data. Arch Neurol 1994;51:916–21.

[38] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations; 2015. San Diego, CA. Available at: http://arxiv.org/abs/1409.1556.

[39] Billones CD, Demetria OJLD, Hostallero DED, Naval PC. DemNet: A Convolutional Neural Network for the Detection of Alzheimer's Disease and Mild Cognitive Impairment. Proceedings of the 2016 Ieee Region 10 Conference (Tencon); 2016. p. 3724–7.

[40] De Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Ann Operations Res 2005;134:19–67.

[41] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

[42] Kolachalama VB, Singh P, Lin CQ, Belghasem ME, Henderson JM, Francis JM, et al. Association of pathological fibrosis with renal survival using deep neural networks. Kidney Int Rep 2018;3:464–75.

[43] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Machine Learn Res 2014;15:1929–58.

[44] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. PMLR 2015;37:448–56.

[45] Chapman KR, Bing-Canar H, Alosco ML, Steinberg EG, Martin B, Chaisson C, et al. Mini Mental State Examination and Logical Memory scores for entry into Alzheimer's disease trials. Alzheimers Res Ther 2016;8:9.

[46] Mckhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's

Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:263–9.

[47] Albert MS, Dekosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:270–9.

[48] de Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis. Med Image Anal 2016;33:94–7.

[49] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

[50] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017;19:221–48.

[51] Nettiksimmons J, Decarli C, Landau S, Beckett L, Alzheimer's Disease Neuroimaging I. Biological heterogeneity in ADNI amnestic mild cognitive impairment. Alzheimers Dement 2014; 10:511–521.e1.

Appendix

Supplementary Table 1
Modified VGG-11 architecture used for model development

Layers	Output channel (k-s-p)	Dropout rate
Conv1_1	64 (3 – 1 - 1)	
Batch Norm, Relu		
Max Pooling	64 (2 - 2 - 0)	
Dropout		0.2
Conv2_1	128 (3 – 1 - 1)	
Batch Norm, Relu		
Max Pooling	128 (2 - 2 - 0)	
Dropout		0.2
Conv3_1	256 (3 – 1 - 1)	
Conv3_2	256 (3 – 1 - 1)	
Batch Norm, Relu		
Max Pooling	256 (2 - 2 - 0)	
Dropout		0.2
Conv4_1	512 (3 – 1 - 1)	
Conv4_2	512 (3 – 1 - 1)	
Batch Norm, Relu		
Max Pooling	512 (2 - 2 - 0)	
Dropout		0.4
Conv5_1	512 (3 – 1 - 1)	
Conv5_2	512 (3 – 1 - 1)	
Batch Norm, Relu		
Max Pooling	512 (2 - 2 - 0)	
Dropout		0.4
Fc_6, Relu	4096	
Dropout		0.5
Fc_7, Relu	4096	
Dropout		0.5
Fc_8, Relu	1000	
Dropout		0.5
Fc_9, Relu	20	
Dropout		0.5
Fc_10, Softmax	2	