Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI

Ethan Ocasio [1], Tim Q Duong [Corresp. 1]

[1] Department of Radiology, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York, United States

Corresponding Author: Tim Q Duong
Email address: tim.duong@einsteinmed.org

Background. While there is no cure for Alzheimer's disease (AD), early diagnosis and accurate prognosis of AD may enable or encourage lifestyle changes, neurocognitive enrichment, and interventions to slow the rate of cognitive decline. The goal of our study was to develop and evaluate a novel deep learning algorithm to predict mild cognitive impairment (MCI) to AD conversion at three years after diagnosis using longitudinal and whole-brain 3D MRI.

Methods. This retrospective study consisted of 320 normal cognition (NC), 554 MCI, and 237 AD patients. Longitudinal data include T1-weighted 3D MRI obtained at initial presentation with diagnosis of MCI and at 12-month follow up. Whole-brain 3D MRI volumes were used without a priori segmentation of regional structural volumes or cortical thicknesses. MRIs of the AD and NC cohort were used to train a deep learning classification model to obtain weights to be applied via transfer learning for prediction of MCI patient conversion to AD at three years post-diagnosis. Two (zero-shot and fine tuning) transfer learning methods were evaluated. Three different convolutional neural network (CNN) architectures (sequential, residual bottleneck, and wide residual) were compared. Data were split into 75% and 25% for training and testing, respectively, with 4-fold cross validation. Prediction accuracy was evaluated using balanced accuracy. Heatmaps were generated.

Results. The sequential convolutional approach yielded slightly better performance than the residual-based architecture, the zero-shot transfer learning approach yielded better performance than fine tuning, and CNN using longitudinal data performed better than CNN using a single timepoint MRI in predicting MCI conversion to AD. The best CNN model for predicting MCI conversion to AD at three years after diagnosis yielded a balanced accuracy of 0.793. Heatmaps of the prediction model showed regions most relevant to the network including the lateral ventricles, periventricular white matter and cortical gray matter.

Conclusions. This is the first convolutional neural network model using longitudinal and whole-brain 3D MRIs without extracting regional brain volumes or cortical thicknesses to predict future MCI to AD conversion at 3 years after diagnosis. This approach could lead to early prediction of patients who are likely to progress to AD and thus may lead to better management of the diseases.

1 **Deep Learning Prediction of Mild Cognitive**
2 **Impairment Conversion to Alzheimer's Disease at 3**
3 **years after diagnosis using Longitudinal and Whole-**
4 **Brain 3D MRI**

6

8 Ethan Ocasio[1], Tim Q Duong[1]
9
10 [1]Department of Radiology, Montefiore Medical Center, Albert Einstein College of Medicine,
11 Bronx, New York
12
13 Corresponding Author:
14 Tim Duong, PhD[1]
15 Department of Radiology, Albert Einstein College of Medicine, 111 E 210th St, Bronx, New
16 York 10467, Email address: Tim.duong@einsteinmed.org

Abstract

21   **Background.** While there is no cure for Alzheimer's disease (AD), early diagnosis and accurate
22   prognosis of AD may enable or encourage lifestyle changes, neurocognitive enrichment, and
23   interventions to slow the rate of cognitive decline. The goal of our study was to develop and
24   evaluate a novel deep learning algorithm to predict mild cognitive impairment (MCI) to AD
25   conversion at three years after diagnosis using longitudinal and whole-brain 3D MRI.

27   **Methods.** This retrospective study consisted of 320 normal cognition (NC), 554 MCI, and 237
28   AD patients. Longitudinal data include T1-weighted 3D MRI obtained at initial presentation with
29   diagnosis of MCI and at 12-month follow up. Whole-brain 3D MRI volumes were used without a
30   priori segmentation of regional structural volumes or cortical thicknesses. MRIs of the AD and
31   NC cohort were used to train a deep learning classification model to obtain weights to be applied
32   via transfer learning for prediction of MCI patient conversion to AD at three years post-
33   diagnosis. Two (zero-shot and fine tuning) transfer learning methods were evaluated. Three
34   different convolutional neural network (CNN) architectures (sequential, residual bottleneck, and
35   wide residual) were compared. Data were split into 75% and 25% for training and testing,
36   respectively, with 4-fold cross validation. Prediction accuracy was evaluated using balanced
37   accuracy. Heatmaps were generated.

38

39   **Results.** The sequential convolutional approach yielded slightly better performance than the
40   residual-based architecture, the zero-shot transfer learning approach yielded better performance
41   than fine tuning, and CNN using longitudinal data performed better than CNN using a single
42   timepoint MRI in predicting MCI conversion to AD. The best CNN model for predicting MCI
43   conversion to AD at three years after diagnosis yielded a balanced accuracy of 0.793. Heatmaps
44   of the prediction model showed regions most relevant to the network including the lateral
45   ventricles, periventricular white matter and cortical gray matter.

46

47   **Conclusions.** This is the first convolutional neural network model using longitudinal and whole-
48   brain 3D MRIs without extracting regional brain volumes or cortical thicknesses to predict future
49   MCI to AD conversion at 3 years after diagnosis. This approach could lead to early prediction of
50   patients who are likely to progress to AD and thus may lead to better management of the
51   diseases.

52

53

54

Background

Alzheimer's disease (AD) is a progressive neurodegenerative disease characterized by loss of memory and other cognitive functions (McKhann et al. 2011). Mild Cognitive Impairment (MCI) is considered a transitional state between normal aging and dementia. Many patients progress from MCI to AD, but others remain stable without developing AD. Although diagnoses of MCI and AD are typically made using neuropsychological tests (Petersen et al. 1999; Jak et al. 2009), imaging methods are also used to diagnose AD and to monitor disease progression because they provide neural correlates of the underlying brain dysfunction in a longitudinal non-invasive manner (Johnson et al. 2012). While there is no cure for AD, early diagnosis and accurate prognosis may enable or encourage lifestyle changes, neurocognitive enrichment, and therapeutic interventions that strive to improve symptoms, or at least slow down mental deterioration, thereby improving the quality of life (Epperly et al. 2017).

Machine learning (ML) is increasingly being used in medicine from disease classification to prediction of clinical progression (de Bruijne 2016; Erickson et al. 2017). ML uses algorithms to learn the relationship amongst different data elements to inform outcomes. Neural networks, a form of ML, are made up of a collection of connected nodes that model the neurons present in a human brain (Graupe 2013). Each connection, similar to a synapse, transmits and receives signals to other nodes. Each node and the connections it forms are initialized with weights which are adjusted throughout training and create mathematical relationships between the input data and the outcomes. In contrast to traditional analysis methods such as logistic regression, neural networks do not require relationships between different input variables and the outcomes to be explicitly specified a priori. In radiology, ML can accurately detect lung nodules on chest X-rays (Harris et al. 2019). In cardiology, ML can detect abnormal EKG patterns (Johnson et al. 2018). ML has also been used to estimate risk, such as in the Framingham Risk Score for coronary heart disease (Alaa et al. 2019), and to guide antithrombotic therapy in atrial fibrillation (Lip et al. 2010) and defibrillator implantation in hypertrophic cardiomyopathy (O'Mahony et al. 2014). Convolutional neural networks (CNNs), a deep-learning method, are widely used for image analysis and analysis of complex data (Lecun et al. 1998; Krizhevsky et al. 2012; Simonyan & Zisserman 2014).

Deep learning classification amongst normal cognition (NC), MCI and AD based on magnetic resonance imaging (MRI) data have been reported (Cheng et al. 2017; Korolev et al. 2017; Wen et al. 2020). By contrast, there are comparatively fewer studies that reported prediction of MCI to AD conversion using deep learning of MRI data (Lian et al. 2018; Lin et al. 2018; Liu et al. 2018; Shmulev & Belyaev 2018; Basaia et al. 2019; Wen et al. 2020). A few ML studies used extracted brain structures or cortical thicknesses, and some used 3D patches from predetermined locations across the brain, but not whole-brain MRI data, to predict MCI to AD conversion (Lian et al. 2018; Liu et al. 2018; Wen et al. 2020). Most of the few prediction studies used single timepoint MRI data. To our knowledge there are only two studies that predicted disease

96   progression using longitudinal imaging data. Bhagwat used a neural network (albeit not deep
97   learning) and extracted cortical thicknesses from MRIs at two time points to predict decline in
98   Mini-Mental Status Exam (MMSE) scores (Bhagwat et al. 2018). Ostertag et al. used a CNN
99   model on whole-brain MRI at two time points to predict decline in MMSE score but did not test
100 their model on an independent testing dataset (Ostertag et al. 2019). These two studies mixed
101 NC, MCI and AD participants and thus accuracies are not applicable to prediction of MCI to AD
102 conversion. *To our knowledge, there are no published studies to date on deep learning to predict*
103 *MCI to AD conversion using longitudinal and whole-brain 3D MRI.*
104
105 The goal of our study was thus to develop and evaluate a novel deep-learning algorithm to
106 predict MCI to AD conversion at three years after diagnosis using longitudinal and whole-brain
107 3D MRI. Longitudinal data include MRI obtained at initial presentation with diagnosis of MCI
108 and at 12-month follow up. Whole-brain 3D MRI volumes were used without a priori
109 segmentation of regional structural volumes or cortical thicknesses. Several convolutional model
110 architectures, transfer learning methods, and methods of merging longitudinal whole-brain 3D
111 MRI data were evaluated to derive the final optimal deep-learning predictive model.
112

Methods

114 **Figure 1** shows the overall design of the experiment. 3D MRIs of the AD and NC cohort were
115 trained in a CNN classification model to obtain weights for transfer learning to be used in the
116 CNN prediction of MCI patient conversion to AD in 3 years after diagnosis. Two (zero-shot and
117 fine tuning) transfer learning methods for prediction were evaluated (Pan & Yang 2010). The
118 zero-shot transfer method used the intact weights obtained from the NC-AD classification
119 without any additional training. The fine-tuning transfer method kept the weights in the
120 convolutional layers frozen while allowing the remaining fully connected layers to change during
121 additional training against the MCI images.
122

Participants

124 Data used in this study was obtained from the Alzheimer's Disease Neuroimaging Initiative
125 (ADNI) database (adni.loni.usc.edu). Patients were taken from the ADNI1, ADNIGO, ADNI2,
126 and ADNI3 patient sets. For the prediction task, inclusion criteria were patients diagnosed with
127 MCI at baseline with MRI taken at baseline and ~12 months after baseline, and with a final
128 diagnosis at 3 years post baseline of either MCI (labeled as stable MCI or sMCI) or AD (labeled
129 as progressive MCI or pMCI). Patients who converted from MCI to AD before their 12-month
130 follow up image were excluded since the analysis of their longitudinal image at that point would
131 represent a diagnostic classification not a prediction. **Table 1** summarizes the participant
132 demographics. Data were split into 75% and 25% for training/validation and testing,
133 respectively, with the training/validation set composed of 415 patients (249 sMCI and 166
134 pMCI), and the testing set composed of 139 patients (84 sMCI and 55 pMCI). Then, we
135 optimized the networks using a 4-fold cross-validation on the training/validation set, resulting in

136   56.25% for training and 18.75% for validation from the complete data in each fold split. In the
137   end, we had four trained models for each experiment configuration, and reported the mean and
138   standard deviation (SD) BA.
139
140   For the AD-NC classification to obtain weights for transfer learning, a separate group of
141   participants with a baseline diagnosis of AD or NC with images taken at baseline and ~12
142   months from baseline were selected. Data for training/validation and testing were randomly split
143   up as 70% and 30% respectively. The training/validation set consisted of 387 patients (160 AD
144   and 227 NC), while the independent testing set consisted of 170 patients (77 AD and 93 NC).
145   Using 4-fold cross validation on the training/validation set resulted in 52.5% for training and
146   17.5% for validation from the complete data in each fold split. The networks were trained to
147   classify the patients between AD and NC against the ground truth diagnosis of each patient using
148   ADNI criteria. For all experiments the study size represents the total available patients in ADNI
149   database fully meeting all inclusion criteria.
150
151   Preprocessing
152   3D volumes of T1-weighted MRI were used as input to the networks. To remove intensity
153   inhomogeneity from the image inputs, T1-weighted images with nonparametric non-uniform
154   intensity normalization (N3) correction (Sled et al. 1998) were selected from the ADNI database.
155   All MRIs were skull stripped with DeepBrain (Itzcovich 2018), then linearly registered against a
156   2mm standard brain with nine degrees of freedom (translation, rotation, scaling) using FSL
157   FLIRT (Jenkinson et al. 2012), and finally min/max intensity normalized. Resulting images had a
158   resolution of 91x109x91 voxels. During training, data augmentation was performed on the
159   training set by rotating each MRI by up to 5% in any direction and randomly flipping them left to
160   right along the sagittal axis.
161
162   Training, validation, and testing
163   Images were split into testing, validation, and testing sets at the patient level in order to avoid
164   data leakage. For both classification and prediction tasks, after assigning labels (either AD vs NC
165   or sMCI vs pMCI), SciPy's train_test_split function was used to randomly split
166   training/validation and testing and cross validation (Virtanen et al. 2020). Stratification was done
167   based on the labels in order to maintain a consistent distribution of diagnoses across the training,
168   validation, and testing datasets. Randomization seed was set up as a constant to ensure the same
169   train/validation/test split was obtained for each experiment run. Balanced accuracy (BA), defined
170   as the average of sensitivity and specificity, was used as the main binary classifier metric to
171   eliminate the inflated accuracy effect caused by imbalanced data sets. For purposes of computing
172   accuracy, we use a standard value of 0.5 as the threshold between the two labels (either NC &
173   AD or pMCI & sMCI). We also computed the area under the receiver operating characteristic
174   curve (AUC) for each run, since it provides a more general measure of the potential performance
175   of a network across a range of thresholds. To prevent data leakage, once the test partition was

176   randomly selected, the test set images were set aside and not used for any training or validation.
177   We repeated each training experiment 4 times, using the standard k-fold cross validation
178   approach where the training/validation set is partitioned into four subsets, using each subset once
179   for validation, and reported both the mean and standard deviation of the BAs for each
180   experiment. This ensures that each image is included at least once in both the validation and
181   training sets, minimizing the potential selection bias that a single random data split may
182   introduce. For each cross-validation fold, classification task training was performed for 200
183   epochs with early stopping based on no improvement in loss function for 80 epochs. Fine tuning
184   after freezing all convolutional layers was performed for 100 epochs with early stopping patience
185   of 40. The weights of the epoch ending with the lowest loss were saved and used to obtain the
186   validation BA, and then the network was run against the test set for the test BA. Additional
187   attempts at fine tuning by also unfreezing the last convolutional block were noted to degrade
188   accuracy, so this approach was not considered any further.
189
190   CNN Architecture
191   The neural network models (**Figure 2**) consisted of a convolutional section followed by a fully
192   connected section (head). Three (sequential, residual with bottleneck, and wide residual) types of
193   convolutional blocks (**Figure 3**) and three head architectures (**Figure 4**) were examined. In
194   addition, we also evaluated these networks using a single (baseline) timepoint MRI and MRIs at
195   two time points (baseline and 12 months). For the dual timepoint networks, three types of
196   networks were explored to incorporate longitudinal data: (1) Siamese network (two identical
197   parallel channels with weights tied together) using a subtraction layer as the merging function,
198   (2) Siamese network with a concatenation merge layer, and (3) a twin network (identical
199   channels with weights independently optimized). Since the flattened set of post-convolution
200   features in the twin architecture is different in each channel, as they are the result of different
201   parameters, there is no rationale for directly subtracting them, so we only considered a
202   concatenation merge option for the twin architecture. For all networks, the final binary classifier
203   layer was fully connected with sigmoid activation. When performing the prediction tasks (using
204   both zero-shot learning and fine-tuning) in the single timepoint experiments, we attempted the
205   same task using both the initial baseline MRI as well as the 1-year follow up image. Using either
206   the single timepoint with the 1-year image or both timepoints together longitudinally represents,
207   for patients who have had MCI for one year and not yet progressed to AD, a prediction of
208   whether they will eventually convert to AD within two more years.
209
210   After initial experimentation, an optimal set of blocks was identified for the sequential and wide
211   residual network styles, namely using 6 blocks with widths (number of activation maps) = {64,
212   128, 256, 512, 1024, 2048}. In the case of the sequential convolution network, each block
213   reduced the resolution via maximum pooling as the width increased, down to 1x1x1 in the final
214   convolutional block. In the case of the wide residual network, convolutions with the use of
215   strides gradually reduced the resolution down to 2x2x2, with a global maximum pooling layer in

216   the head portion of the network. When a convolutional layer processes an input whose size is
217   odd-numbered in any of its dimensions (2n-1 for any integer n) the resulting output of a stride 2
218   convolution with zero-padding will be of size n for the corresponding dimension. For example,
219   since the input image has resolution 91x109x91, the output after the first stride 2 convolution
220   will be 46x55x46. For the bottleneck residual, the final convolutional resolution was also 2x2x2,
221   achieved via strides, but this required 7 blocks with widths = {64, 64, 128, 256, 512, 1024,
222   2048}. The bottleneck architecture used 1x1x1 convolutions for resolution reduction, so the
223   behavior was slightly different when the resolution had odd numbers, allowing for 7 instead of 6
224   blocks until the resolution was down to 2x2x2. The portion of the network with flattened non-
225   convolutional fully connected layers after the last convolutional layer up until the final binary
226   classifier is known as the "head." After initial analysis of networks with varying heads, a global
227   maximum pooling operation resulting in a fully connected layer with a number of nodes equal to
228   the number of activation maps (width) of the last convolutional step, followed directly by a
229   single final dense prediction layer, was selected as the optimal fully connected layer architecture
230   (**Figure 4A**).

231

232   Training was initially attempted using both non-adaptive (SGD with Nesterov momentum), as
233   well as adaptive (Adam) optimizers (Kingma & Ba 2019). The Adam optimizer was able to
234   achieve reductions in loss function with accompanying increase in accuracy much more rapidly
235   and aggressively. However, without a scheduled reduction of the base learning rate, the network
236   became unstable in the latter epochs with rapid swings in the loss function. The use of an
237   exponentially decaying learning rate schedule consistently stabilized both the loss and accuracy
238   curves in an optimal fashion. The final selected optimization approach was thus the Adam
239   optimizer with an exponentially decreasing learning rate schedule with expected initial rate L
240   R_{start} and final rate LR_{end} where t is the current epoch and T is the final expected number of
241   epochs:

242

243    $$LR_{epoch} = LR_{start}(\frac{LR_{end}}{LR_{start}})^{t/T}$$

244

245   L2 regularization as also added in the convolutional layers in all networks. For the sequential
246   model, we used a regularization parameter of 0.005, and for the residual models we used a
247   parameter of 0.0001. All training was performed using Tensorflow2/Keras python library, on
248   Google Compute Platform virtual instances with Tesla V-100 GPU acceleration.

249

250   Network visualization by heatmap
251   To visualize the brain regions that are most relevant to the network, the Grad-CAM (Selvaraju et
252   al. 2017) technique was modified to work in 3 dimensions for generating heatmaps. Since the
253   models reduced the resolution of the image information within the convolution blocks down to
254   2x2x2 voxels or less, the 3D Grad-CAM technique was applied to higher convolutional layers

255 (with resolution close to 40 voxels per axis) to obtain more useful visualization heatmaps. This
256 approach enabled visual highlighting of the sections of the images that were most significant to
257 the network. Heatmaps were obtained during the execution of the prediction models.
258

259 **Results**

260

261 AD versus NC classification to generate weights
262 **Figure 5** shows training curves for sequential single channel and wide residual dual channel
263 training for the classification experiments. Overall, loss functions converged and leveled off at
264 around 75 epochs. Other models showed similar convergence characteristics. **Table 2** details the
265 results of the classification experiments to generate weights. For the single timepoint networks,
266 sequential architecture performed best (BA = 0.860) followed by wide residual (BA = 0.840) and
267 bottleneck residual (BA=0.727) on the testing data. For the dual timepoint networks, Siamese
268 network with subtraction performed poorly overall with all architectures (BA 0<0.65) and that
269 the twin non-Siamese approach with merge concatenation performed best for dual channels. The
270 wide residual (BA=0.887) performed best followed by sequential (BA=0.876) and bottleneck
271 residual (BA=0.800). Training time for each run was approximately 60-90 minutes. After model
272 was trained, classification of a patient takes two seconds or less (most of this time is loading the
273 images from storage into memory).
274

275 Prediction of AD conversion at 3 years after diagnosis
276 **Figure 6** shows the training curve for one of the dual sequential transfer learning fine-tuning
277 attempts for the prediction experiment. Additional training did not improve the accuracy even
278 though there was some clear reduction in loss function during training. Other models showed
279 similar characteristics. **Table 3** summarizes the results of the prediction experiments, showing
280 the outcomes of the dual and single timepoint networks using both zero-shot and fine-tuning
281 transfer learning. The bottleneck residual convolutional style, because it performed much worse
282 for classification, was not considered for prediction. For the single timepoint experiments, the
283 12-month images performed better than the initial baseline images during the classification task.
284 The dual timepoint networks performed better than the single timepoints. For zero-shot best
285 results were obtained from the sequential model, with the dual sequential producing a 0.795
286 average BA against the test set, followed by the single timepoint sequential (using the 1-year
287 image as input) with a BA of 0.774. Transfer learning with fine tuning in all cases resulted in
288 *lowered* accuracy as compared with the zero-shot approach. This occurred whether fine tuning
289 was attempted by unfreezing only the fully connected layers or also the last convolutional layer.
290 Training time for each fine-tuning experiment was approximately 15-30 minutes. Prediction of
291 conversion took using a trained model (either zero-shot or fine-tuned) took two seconds or less.
292

293   To visualize the brain regions that are most relevant to ML algorithms, post-training heatmaps
294   for the wide residual dual network were generated (**Figure 7**). The highlighted structures
295   included the lateral ventricles, periventricular white matter, and cortical surface gray matter.
296
297

Discussion

299   This study developed and evaluated a few sophisticated ML algorithms to predict which MCI
300   patients would convert to AD at three years post-diagnosis using longitudinal whole-brain 3D
301   MRI without a priori segmentation of regional structural volumes or cortical thicknesses. MRI
302   data used for prediction were obtained at baseline and one year after baseline. The sequential
303   convolutional approach yielded slightly better performance than the residual-based architecture,
304   the zero-shot transfer learning approach yielded better performance than fine tuning, and the
305   CNN using longitudinal data performed better than the CNN using a single timepoint MRI in
306   predicting MCI conversion to AD. The best CNN model for predicting MCI conversion to AD at
307   3 years after diagnosis yielded a BA of 0.793.
308
309   Our predictive model used whole-brain MRIs without extract regional brain volumes and cortical
310   thicknesses. We also evaluated multiple longitudinal network configurations (i.e., Siamese and
311   non-Siamese twin networks with subtraction and concatenation as the merge function).
312   Longitudinal images were found to be optimally processed by a twin architecture with
313   concatenation merge. The dual timepoint network performed better regardless of whether the
314   initial or the follow up image was used for the single timepoint. Restricting the network in a
315   Siamese configuration where the weights of both channels are identical or using a subtraction
316   merge function resulted in worse prediction, which suggests that the networks take full
317   advantage of the additional information provided by the second time point data when they were
318   allowed to train each channel with separate weights.
319
320   We employed 3D MRI instead of 2D multi-slice MRI. Previous studies have also reported MCI
321   to AD prediction using a sequential full volume 3D architecture have obtained BAs of 0.75
322   (Basaia et al. 2019) and 0.73 (Wen et al. 2020), while a study using residual architecture showed
323   a resulting BA of 0.67 (Shmulev & Belyaev 2018) but did not involve longitudinal MRI data.
324   Some studies used predetermined 3D patches uniformly sampled across the brain (Lian et al.
325   2018; Liu et al. 2018; Wen et al. 2020). A limitation of the 3D-patch approach is that a
326   subsequent fusion of the results via some kind of ensemble or voting method is needed to obtain
327   a subject-level prediction, and brain-wide anatomic relationships are not taken into account.
328   There are two previous related studies that used longitudinal MRI data for prediction of MCI or
329   AD disease progression. Bhagwat et al. employed baseline and 1-year MRIs with Siamese neural
330   network with concatenation merge to predict a pattern of decline in patients' MMSE score,
331   yielding an accuracy of 0.95 (Bhagwat et al. 2018). In contrast to our study, regional cortical
332   thicknesses, a non-convolutional method, and clinical variables were used. The use of clinical

333   variables could have contributed substantially to higher accuracy. Ostertag et al. used a similar
334   Siamese network but employed whole-brain MRI to predict decline in patients' MMSE score,
335   yielding a validation accuracy of 0.90, but no independent evaluation on a separate test dataset
336   was performed (Ostertag et al. 2019). Moreover, these two studies differed from ours in that they
337   mixed AD, NC, and MCI patients together, and thus their prediction accuracies are not directly
338   comparable to those from MCI to AD conversion studies (thus accuracies are not applicable to
339   prediction of MCI to AD conversion) because the baseline diagnosis of NC or AD by itself is a
340   strong predictor of neurocognitive decline.
341
342   The use of a Siamese network architecture to analyze longitudinal changes in disease progression
343   from medical images was explored by Li et al. and specifically studied in AD brain MRIs by
344   Bhagwat et al. and Ostertag et al. (Bhagwat et al. 2018; Ostertag et al. 2019; Li et al. 2020). The
345   idea behind Siamese networks is that both images are processed by the convolutional layers with
346   identical parameters, with equivalent flattened sets of features for each image at the end of the
347   convolutions. Thus, theoretically, a direct subtraction merge of the corresponding flattened
348   features would represent a measure of the progression of the images—presumably an MCI
349   patient whose structural MRI features have worsened in a year would be more likely to progress
350   to AD than a patient whose features remain stable. However, a simple subtraction merge may
351   result in loss of predictive information if there are particular features that are predictive of
352   progression regardless of whether they have changed between baseline and 1-year. Thus, we also
353   explore the concatenation merge. In addition, a twin (non-Siamese) network with separate
354   parameters may provide, partly due to the additional power of doubling the number of
355   convolutional parameters, better predictive capacity, so we also explored this architecture. Since
356   the flattened set of post-convolution features in the twin architecture is different in each channel,
357   as they are the result of different parameters, there is no rationale for directly subtracting them,
358   so we only considered a concatenation merge option for the twin architecture.
359
360   Although for the initial classification task, the twin wide residual network performed best among
361   all architectures, after the transfer learning the twin sequential network was the overall best
362   performer. In the single channel variants, the sequential networks performed best. The bottleneck
363   variant of the residual network performed the worst amongst all architectures. In general, the
364   residual networks provide the benefit of reducing the vanishing gradient problem, as compared
365   with a non-residual sequential style. The bottleneck in particular is meant to strongly prevent
366   vanishing gradients. Since vanishing gradients did not appear significantly during training, the
367   advantages of the residual network appeared not to materialize, and thus, overall, the sequential
368   networks seemed best fit for 3D MRI whole-brain analysis.
369
370   Heatmaps
371   Heatmaps enabled visualization of the brain regions that were most relevant to ML algorithms to
372   predict MCI and AD conversion. The most salient structures on the heatmaps were the lateral

373   ventricles, periventricular deep white matter as well as extensive cortical gray matter. Ventricular
374   enlargement and atrophy are known to be associated with AD. Reduction in white-matter volume
375   has been described in AD, including some the specific regions that our heatmap analysis found to
376   be on interest (Smith et al. 2000; Guo et al. 2010; Kao et al. 2019), including the cingulate gyrus
377   (Brun & Gustafson 1976; Hirono et al. 1998; Jones et al. 2006), the middle occipital gyrus
378   (Zhang & Wang 2015), and the putamen (Pini et al. 2016). Other brain regions that have shown
379   to be associated with development of AD, such as the default node network and hippocampus,
380   are not uniformly highlighted in the heatmaps. Our analysis approach is different from previous
381   analysis and does not specifically identify networks, although amongst the heatmaps shown,
382   there were components that were part of the default mode networks and hippocampus. In other
383   words, our analysis did not specifically test whether hippocampus or default mode networks are
384   predictive of MCI to AD conversion. It is possible that, given our MRI is based on structural
385   changes, hippocampus and default mode networks might not have developed atrophy to be
386   informative to prediction conversion.
387
388
389   Other technical considerations
390   We examined three different convolutional architectures to identify the best performance
391   prediction model. Two residual variants were compared, with the wide residual network
392   performing better than the bottleneck variant, and the non-residual sequential network
393   performing better than both residual types. The two residual approaches compared here were 3D
394   modifications of ResNet (He et al. 2016). The bottleneck variation used pre-activation, a
395   technique where the batch normalization and activation layers precede the convolutions. The
396   term "bottleneck" refers to a design where each residual block includes two initial layers with
397   narrower widths. The second residual variant examined for comparison was the wide residual
398   network (Zagoruyko & Komodakis 2016). In this approach the widths were progressively
399   increased, with an additional dropout layer between two convolutions in each residual block. The
400   sequential model we tested was a 3D extension of the 2D VGG model (Simonyan & Zisserman
401   2014), with sequential blocks formed by a combination of convolutional layers followed by
402   pooling layers.
403
404   We also examined the relative performance of two transfer learning approaches. Zero-shot
405   technique performed better than fine tuning. Further fine-tuning with the sMCI vs pMCI data
406   reduced the accuracy of the prediction network from that obtained via the exclusive use of AD vs
407   NC data for *classification* task training. The lack of training power of the MCI data suggests that
408   brain images with either AD or NC, with their more discriminant anatomic features, are more
409   suited for training a network eventually used for detecting the more subtle distinctions between
410   pMCI and sMCI.
411

412   We also carefully prevented data leakage by splitting the training and testing datasets at patient
413   level, ensuring that no data from the same patient would end up in both groups (Wen et al. 2020).
414   Another type of leakage we avoided occurs when data are used for training the classification are
415   also used for prediction task. Finally, in this study the testing set results were collected only after
416   all training was completed to prevent a third possible kind of leakage, namely where results from
417   the test set influence the selection of hyperparameters or architecture. We also excluded patients
418   who converted from MCI to AD before their 12-month follow up.
419

420   In several cases for both classification and prediction we observed that the BA for the testing
421   dataset had a slightly higher mean and lower standard deviation than the corresponding results
422   for validation. This higher variation in the validation experiments could potentially be explained
423   by the fact that each cross-validation fold has a different validation set of images while all the
424   testing results are obtained from the same single test set applying the different trained models.
425   This higher variation also means that a single lower BA result in one of the validation folds
426   could pull down the mean validation BA.
427

428   Limitations and future directions
429   The increase in BA obtained by using the longitudinal MRI (0.795 vs 0.774) was modest,
430   although both techniques represented an increase as compared to other published predictions of
431   MCI conversion to AD. If the longitudinal MRI is otherwise available, it seems evident that the
432   incremental improvement in predictive accuracy would justify its use. It is unclear, however, that
433   without other reasons for performing a 1-year follow up MRI, this increase in predictive
434   accuracy would represent a new indication from a cost-effectiveness perspective. Thus, a
435   comprehensive cost-benefit model analysis would be useful in this area.
436

437   The study used only anatomical MRI data. Multiparametric MRI (such as diffusion-tensor
438   imaging, task functional MRI and resting-state MRI) will be incorporated into these models in
439   the future. Similarly, other modalities such at Positron Emission Tomography (PET) and non-
440   imaging clinical data can also be included in the model. Further studies will need to apply this
441   approach to other datasets to improve generalizability. Future studies should investigate MCI to
442   AD conversion at 1, 2 and 5 years post-diagnosis.
443

444   Our model is a predictive model approach that employs machine learning based on whole-brain
445   anatomical MRI to predict MCI to AD conversion. Future studies will need to compare different
446   predictive models including those that predict MCI to AD conversion based on extracted volume
447   and cortical thickness as obtained using tools such as FastSurfer (Henschel et al. 2020). To do so,
448   we will first systematically explore various methods to extract volume and cortical thickness,
449   explore various approaches (such neural networks and support vector machines) to predict MCI
450   to ADC conversion, and use these methods to do head-to-head comparisons on the same datasets.
451

452 Deep survival analysis (Ranganath et al. 2016) has been applied to the prediction of conversion
453 to AD. Nakagawa et al. used deep survival analysis to model the prediction of conversion from
454 either MCI or NC subjects to AD using volumetric data from MRI (Nakagawa et al. 2020). A
455 future extension of this analysis should investigate the use of data from the CNN models, both
456 single-channel and longitudinal, using features extracted at the end of the convolutional layers.
457
458
459

460 ## Conclusions
461 This is the first convolutional neural network model using longitudinal and whole-brain 3D
462 MRIs without extracting regional brain volumes or cortical thicknesses to predict future MCI to
463 AD conversion. This framework set the stage for further studies of additional data time points,
464 different image types, and non-image data to further improve prediction accuracy of MCI to AD
465 conversion. Accurate prognosis could lead to better management of the diseases, thereby
466 improving the quality of life.
467
468

469 ## References
470

471 Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, and van der Schaar M. 2019. Cardiovascular
472 disease risk prediction using automated machine learning: A prospective study of
473 423,604 UK Biobank participants. *PLoS One* 14:e0213653.
474 10.1371/journal.pone.0213653
475 Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, Filippi M, and Alzheimer's
476 Disease Neuroimaging I. 2019. Automated classification of Alzheimer's disease and mild
477 cognitive impairment using a single MRI and deep neural networks. *Neuroimage Clin*
478 21:101645. 10.1016/j.nicl.2018.101645
479 Bhagwat N, Viviano JD, Voineskos AN, Chakravarty MM, and Alzheimer's Disease
480 Neuroimaging I. 2018. Modeling and prediction of clinical symptom trajectories in
481 Alzheimer's disease using longitudinal data. *PLoS Comput Biol* 14:e1006376.
482 10.1371/journal.pcbi.1006376
483 Brun A, and Gustafson L. 1976. Distribution of cerebral degeneration in Alzheimer's disease. A
484 clinico-pathological study. *Arch Psychiatr Nervenkr (1970)* 223:15-33.
485 10.1007/BF00367450
486 Cheng D, Liu M, Fu J, and Wang Y. 2017. Classification of MR brain images by combination of
487 multi-CNNs for AD diagnosis. Ninth International Conference on Digital Image
488 Processing (ICDIP 2017): SPIE.
489 de Bruijne M. 2016. Machine learning approaches in medical image analysis: From detection to
490 diagnosis. *Med Image Anal* 33:94-97. 10.1016/j.media.2016.06.032
491 Epperly T, Dunay MA, and Boice JL. 2017. Alzheimer Disease: Pharmacologic and
492 Nonpharmacologic Therapies for Cognitive and Functional Symptoms. *Am Fam*
493 *Physician* 95:771-778.

494 Erickson BJ, Korfiatis P, Akkus Z, and Kline TL. 2017. Machine Learning for Medical Imaging.
495   *Radiographics* 37:505-515. 10.1148/rg.2017160130
496 Graupe D. 2013. *Principles of artificial neural networks*. New Jersey: World Scientific.
497 Guo X, Wang Z, Li K, Li Z, Qi Z, Jin Z, Yao L, and Chen K. 2010. Voxel-based assessment of
498   gray and white matter volumes in Alzheimer's disease. *Neurosci Lett* 468:146-150.
499   10.1016/j.neulet.2009.10.086
500 Harris M, Qi A, Jeagal L, Torabi N, Menzies D, Korobitsyn A, Pai M, Nathavitharana RR, and
501   Ahmad Khan F. 2019. A systematic review of the diagnostic accuracy of artificial
502   intelligence-based computer programs to analyze chest x-rays for pulmonary
503   tuberculosis. *PLoS One* 14:e0221339. 10.1371/journal.pone.0221339
504 He K, Zhang X, Ren S, and Sun J. 2016. Deep residual learning for image recognition.
505   Proceedings of the IEEE conference on computer vision and pattern recognition. p 770-
506   778.
507 Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, and Reuter M. 2020. Fastsurfer-a fast and
508   accurate deep learning based neuroimaging pipeline. *Neuroimage* 219:117012.
509 Hirono N, Mori E, Ishii K, Ikejiri Y, Imamura T, Shimomura T, Hashimoto M, Yamashita H,
510   and Sasaki M. 1998. Hypofunction in the posterior cingulate gyrus correlates with
511   disorientation for time and place in Alzheimer's disease. *J Neurol Neurosurg Psychiatry*
512   64:552-554. 10.1136/jnnp.64.4.552
513 Itzcovich I. 2018. DeepBrain Extractor. GitHub.
514 Jak AJ, Bondi MW, Delano-Wood L, Wierenga C, Corey-Bloom J, Salmon DP, and Delis DC.
515   2009. Quantification of five neuropsychological approaches to defining mild cognitive
516   impairment. *Am J Geriatr Psychiatry* 17:368-375. 10.1097/JGP.0b013e31819431d5
517 Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, and Smith SM. 2012. Fsl. *Neuroimage*
518   62:782-790. 10.1016/j.neuroimage.2011.09.015
519 Johnson KA, Fox NC, Sperling RA, and Klunk WE. 2012. Brain imaging in Alzheimer disease.
520   *Cold Spring Harb Perspect Med* 2:a006213. 10.1101/cshperspect.a006213
521 Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, and Dudley
522   JT. 2018. Artificial Intelligence in Cardiology. *J Am Coll Cardiol* 71:2668-2679.
523   10.1016/j.jacc.2018.03.521
524 Jones BF, Barnes J, Uylings HB, Fox NC, Frost C, Witter MP, and Scheltens P. 2006.
525   Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric
526   MRI study. *Cereb Cortex* 16:1701-1708. 10.1093/cercor/bhj105
527 Kao YH, Chou MC, Chen CH, and Yang YH. 2019. White Matter Changes in Patients with
528   Alzheimer's Disease and Associated Factors. *J Clin Med* 8. 10.3390/jcm8020167
529 Kingma DP, and Ba JA. 2019. A method for stochastic optimization. arXiv 2014. *arXiv preprint*
530   *arXiv:14126980* 434.
531 Korolev S, Safiullin A, Belyaev M, and Dodonova Y. 2017. Residual and plain convolutional
532   neural networks for 3D brain MRI classification. 2017 IEEE 14th International
533   Symposium on Biomedical Imaging (ISBI 2017): IEEE. p 835-838.
534 Krizhevsky A, Sutskever I, and Hinton GE. 2012. Imagenet classification with deep
535   convolutional neural networks. Advances in neural information processing systems. p
536   1097-1105.
537 Lecun Y, Bottou L, Bengio Y, and Haffner P. 1998. Gradient-based learning applied to
538   document recognition. *Proceedings of the IEEE* 86:2278-2324. 10.1109/5.726791

539   Li MD, Chang K, Bearce B, Chang CY, Huang AJ, Campbell JP, Brown JM, Singh P, Hoebel
540   KV, and Erdoğmuş D. 2020. Siamese neural networks for continuous disease severity
541   evaluation and change detection in medical imaging. *NPJ digital medicine* 3:1-9.
542   Lian C, Liu M, Zhang J, and Shen D. 2018. Hierarchical Fully Convolutional Network for Joint
543   Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI. *IEEE*
544   *Trans Pattern Anal Mach Intell* 42:880-893. 10.1109/TPAMI.2018.2889096
545   Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, Guo G, Xiao M, Du M, Qu X, and TAsDNI.
546   2018. Convolutional Neural Networks-Based MRI Image Analysis for the Alzheimer's
547   Disease Prediction From Mild Cognitive Impairment. *Frontiers in Neuroscience* 12.
548   10.3389/fnins.2018.00777
549   Lip GY, Nieuwlaat R, Pisters R, Lane DA, and Crijns HJ. 2010. Refining clinical risk
550   stratification for predicting stroke and thromboembolism in atrial fibrillation using a
551   novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*
552   137:263-272. 10.1378/chest.09-1584
553   Liu M, Zhang J, Adeli E, and Shen D. 2018. Landmark-based deep multi-instance learning for
554   brain disease diagnosis. *Med Image Anal* 43:157-168. 10.1016/j.media.2017.10.005
555   McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Jr., Kawas CH, Klunk WE,
556   Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P,
557   Carrillo MC, Thies B, Weintraub S, and Phelps CH. 2011. The diagnosis of dementia due
558   to Alzheimer's disease: recommendations from the National Institute on Aging-
559   Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.
560   *Alzheimers Dement* 7:263-269. 10.1016/j.jalz.2011.03.005
561   Nakagawa T, Ishida M, Naito J, Nagai A, Yamaguchi S, Onoda K, and Initiative AsDN. 2020.
562   Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI
563   images. *Brain communications* 2:fcaa057.
564   O'Mahony C, Jichi F, Pavlou M, Monserrat L, Anastasakis A, Rapezzi C, Biagini E, Gimeno JR,
565   Limongelli G, McKenna WJ, Omar RZ, Elliott PM, and Hypertrophic Cardiomyopathy
566   Outcomes I. 2014. A novel clinical risk prediction model for sudden cardiac death in
567   hypertrophic cardiomyopathy (HCM risk-SCD). *Eur Heart J* 35:2010-2020.
568   10.1093/eurheartj/eht439
569   Ostertag C, Beurton-Aimar M, and Urruty T. 2019. 3D-SiameseNet to analyze brain MRI. 10th
570   International Conference on Pattern Recognition Systems (ICPRS-2019): IET. p 18-23.
571   Pan SJ, and Yang Q. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge*
572   *and Data Engineering* 22:1345-1359. 10.1109/TKDE.2009.191
573   Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, and Kokmen E. 1999. Mild
574   cognitive impairment: clinical characterization and outcome. *Arch Neurol* 56:303-308.
575   10.1001/archneur.56.3.303
576   Pini L, Pievani M, Bocchetta M, Altomare D, Bosco P, Cavedo E, Galluzzi S, Marizzoni M, and
577   Frisoni GB. 2016. Brain atrophy in Alzheimer's Disease and aging. *Ageing Res Rev*
578   30:25-48. 10.1016/j.arr.2016.01.002
579   Ranganath R, Perotte A, Elhadad N, and Blei D. 2016. Deep Survival Analysis. In: Finale D-V,
580   Jim F, David K, Byron W, and Jenna W, editors. Proceedings of the 1st Machine
581   Learning for Healthcare Conference. Proceedings of Machine Learning Research: PMLR.
582   p 101--114.

583   Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. 2017. Grad-cam: Visual
584   explanations from deep networks via gradient-based localization. Proceedings of the
585   IEEE international conference on computer vision. p 618-626.
586   Shmulev Y, and Belyaev M. 2018. Predicting Conversion of Mild Cognitive Impairments to
587   Alzheimer's Disease and Exploring Impact of Neuroimaging. In: Stoyanov D, Taylor Z,
588   Ferrante E, Dalca AV, Martel A, Maier-Hein L, Parisot S, Sotiras A, Papiez B, Sabuncu
589   MR, and Shen L, editors. Graphs in Biomedical Image Analysis and Integrating Medical
590   Imaging and Non-Imaging Modalities. Cham: Springer International Publishing. p 83-91.
591   Simonyan K, and Zisserman A. 2014. Very deep convolutional networks for large-scale image
592   recognition. *arXiv preprint arXiv:14091556.*
593   Sled JG, Zijdenbos AP, and Evans AC. 1998. A nonparametric method for automatic correction
594   of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging* 17:87-97.
595   Smith CD, Snowdon DA, Wang H, and Markesbery WR. 2000. White matter volumes and
596   periventricular white matter hyperintensities in aging and dementia. *Neurology* 54:838-
597   842. 10.1212/wnl.54.4.838
598   Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,
599   Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ,
600   Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore
601   EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris
602   CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, and SciPy C. 2020. SciPy
603   1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17:261-
604   272. 10.1038/s41592-019-0686-2
605   Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-Gonzalez J, Routier A, Bottani S, Dormont D,
606   Durrleman S, Burgos N, Colliot O, Alzheimer's Disease Neuroimaging I, Australian
607   Imaging B, and Lifestyle flagship study of a. 2020. Convolutional neural networks for
608   classification of Alzheimer's disease: Overview and reproducible evaluation. *Med Image*
609   *Anal* 63:101694. 10.1016/j.media.2020.101694
610   Zagoruyko S, and Komodakis N. 2016. Wide residual networks. *arXiv preprint*
611   *arXiv:160507146.*
612   Zhang Y, and Wang S. 2015. Detection of Alzheimer's disease by displacement field and
613   machine learning. *PeerJ* 3:e1251. 10.7717/peerj.1251
614

Figure 1

Overview of experimental design.

AD and NC MRI data were first trained to obtain weights (a classification task) for transfer learning (blue). After training, the weights are transferred to the prediction task (green) to predict whether patients will remain stable or progress within three years. Two different transfer learning methods were studied. With zero-shot, no further training was performed after the transfer, so the MCI images were analyzed for prediction by the network with the same weights copied over from the classification task. With fine-tuning, after weights are copied over from the classification task for initialization, additional training is performed against the MCI image data.

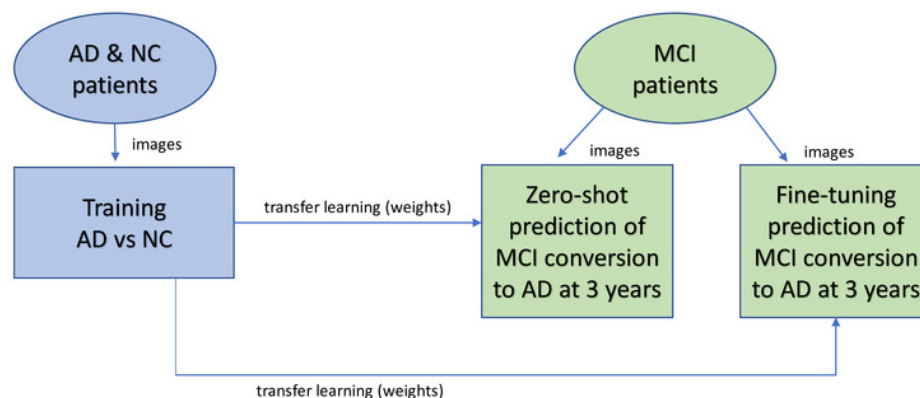



Figure 2

Single and dual time point CNN architecture.

(A) Single timepoint CNN. For classification, input consisted of a single timepoint full-subject 3D MRI of patients diagnosed at baseline as either AD or CN, and output was binary classification of AD vs CN. For prediction, input was a single timepoint full-subject 3D MRI of patients diagnosed as MCI and output was a binary prediction of whether the patient progressed (pMCI) or remained stable (sMCI) 3 years later. **(B)** Dual timepoint CNN. Input included 3D MRI images obtained at both baseline and 12 months, with the patient population and output categories identical than those used for single timepoint for classification and prediction. Both kinds of networks began with a series of convolutional blocks, followed by flattening into one or more fully connected layers ending in a final binary choice of classification or prediction.

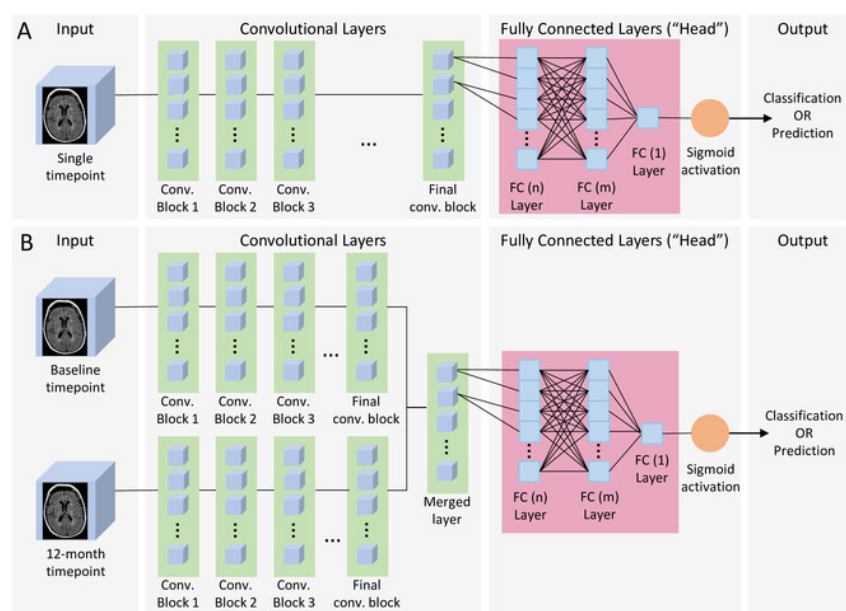



Figure 3

Sequential, residual with bottleneck, and wide residual CNN blocks.

The convolutional layers portion of the network was organized as a series of blocks, each one with an increasing number K of activation maps (width), and with a corresponding decrease in resolution obtained by either pooling or stride during convolution. The figures detail the individual layers that compose a single block. **(A)** Sequential convolutional block. Each block was composed of a single 3x3x3 convolution, followed by batch normalization, ReLU activation, and max pooling to reduce the resolution. **(B)**Residual bottleneck with preactivation convolutional block. Convolutions were preceded by batch normalization and ReLU activation. Two bottleneck 3x3x3 convolutions have a width of K/4 followed by a final 1x1x1 convolution with K width. In parallel the skip residual used a 1x1x1 convolution to match the width and resolution. In this architecture the first residual block was preceded by an initial batch normalization followed by a single 5x5x5 convolution, plus one final batch normalization and ReLU activation after the last block (not shown). **(C)** Wide Residual Network convolutional block. In this architecture the batch normalization and activations occured after the convolutional layers. Each block had two 3x3x3 convolutional layers with 3D spatial dropout in between, plus a 1x1x1 skip residual convolution to match width and resolution.

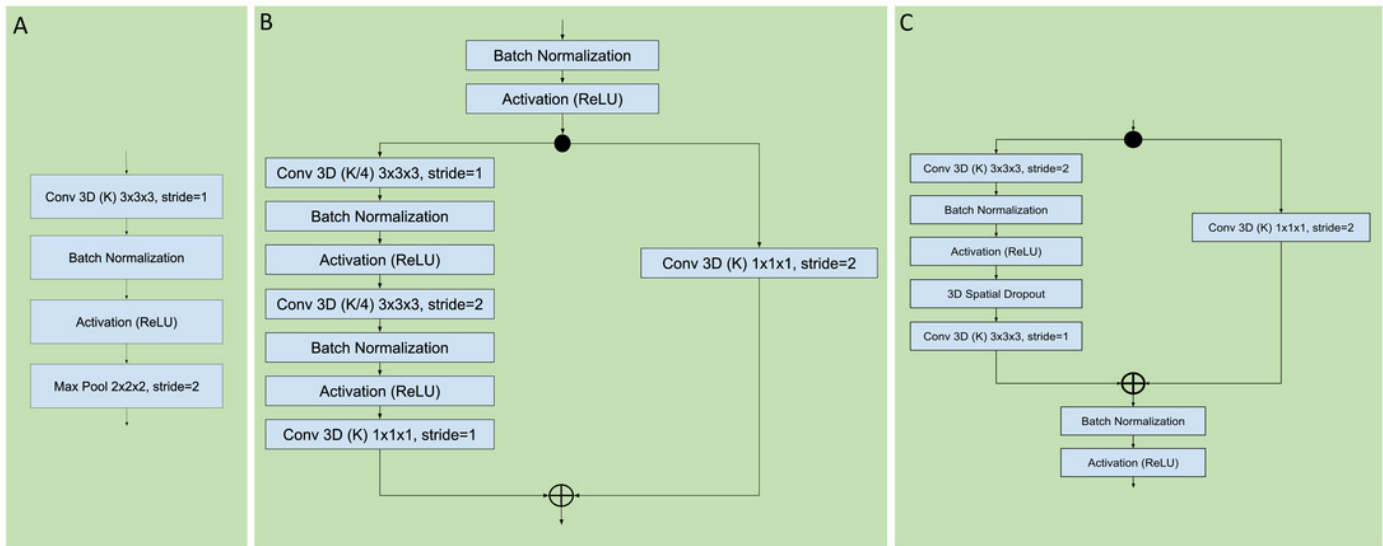



Figure 4

Three head architectures.

(A) 3D global maximum pooling fully connected block. The global pooling inherently flattened the nodes into a fully connected layer with N nodes directly followed by the final binary classifier layer. **(B)** Long fully connected block. After flattening into a layer of N nodes, there are two sets of fully connected (size 2048 and 1024), batch normalization, and leaky ReLU activation layers separated by a single dropout layer, before the final binary classifier. **(C)** Medium fully connected block. Initial 3D max pooling is followed by flattening into a fully connected layer of size N followed by an additional fully connected layer of size 128 and ReLU activation.

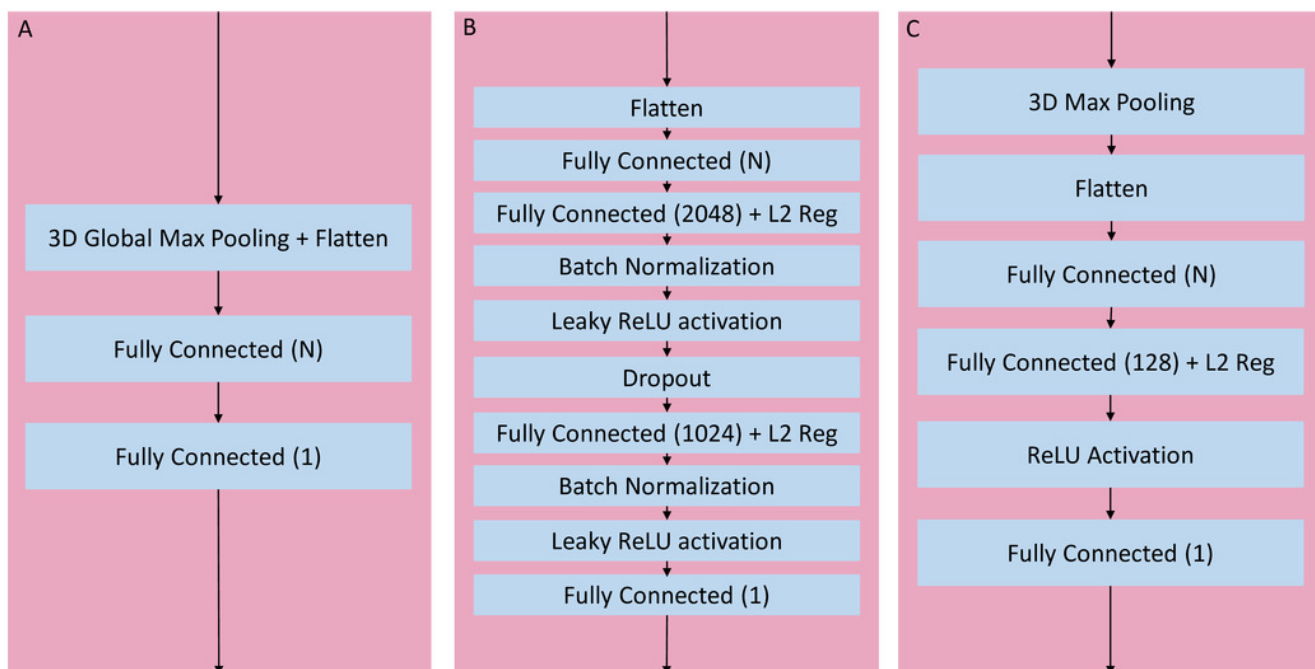



Figure 5

Training curves during classification.

Loss and Accuracy curves during training for both training and validation sets. **For** sequential network and single timepoint, **(A)** loss, **(B)** accuracy. For wide residual network and dual timepoints, **(C)** loss, **(D)** accuracy. Solid lines are smoothed with 0.8 factor and faint lines show the unsmoothed values for each epoch.

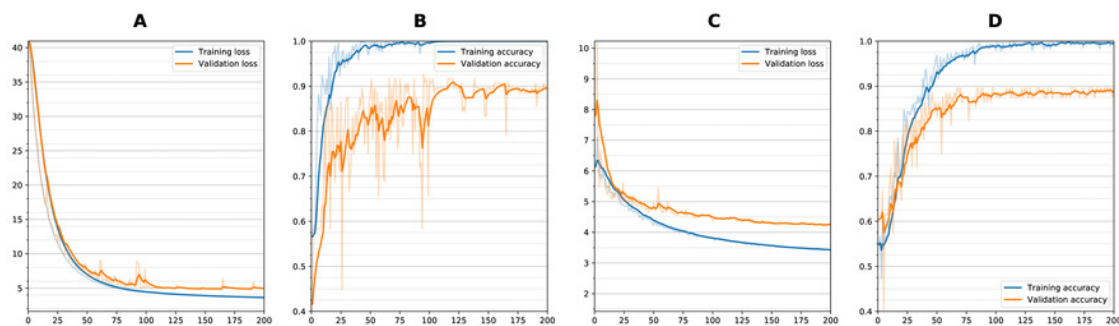



Figure 6

Training curve during fine tuning for prediction.

(A) loss function per epoch and **(B)** accuracy per epoch) during transfer learning fine tuning (sequential dual channel). Weights were initialized after training with AD vs. NC and then frozen at the convolutional layers, then additional training performed with the sMCI vs. pMCI data. There is an initial reduction in loss which stabilizes after 10 epochs, with no increase in accuracy.

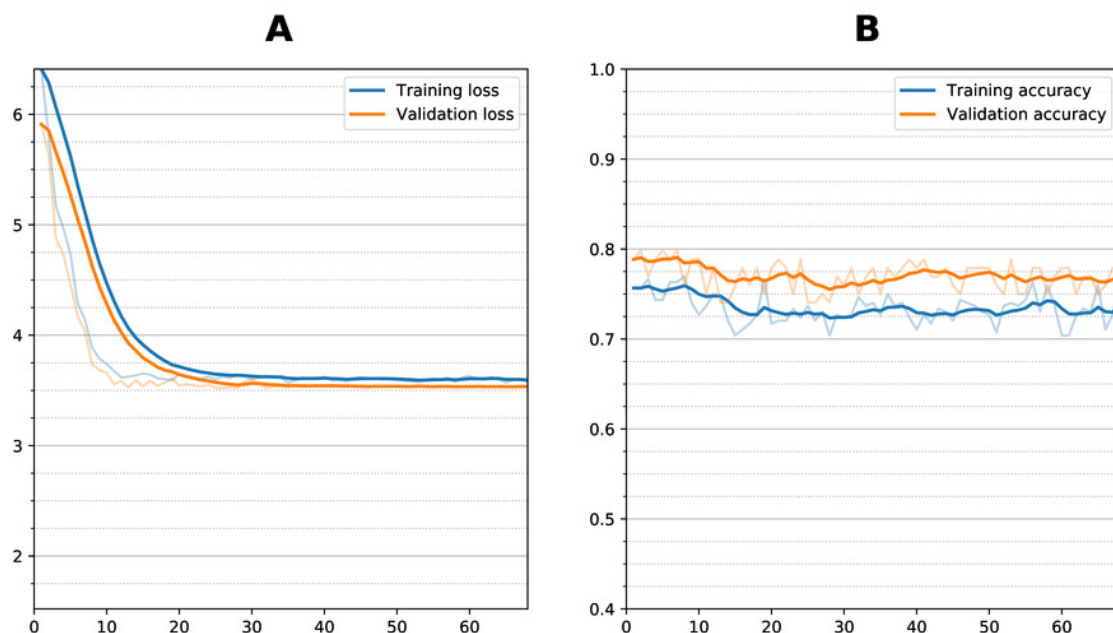
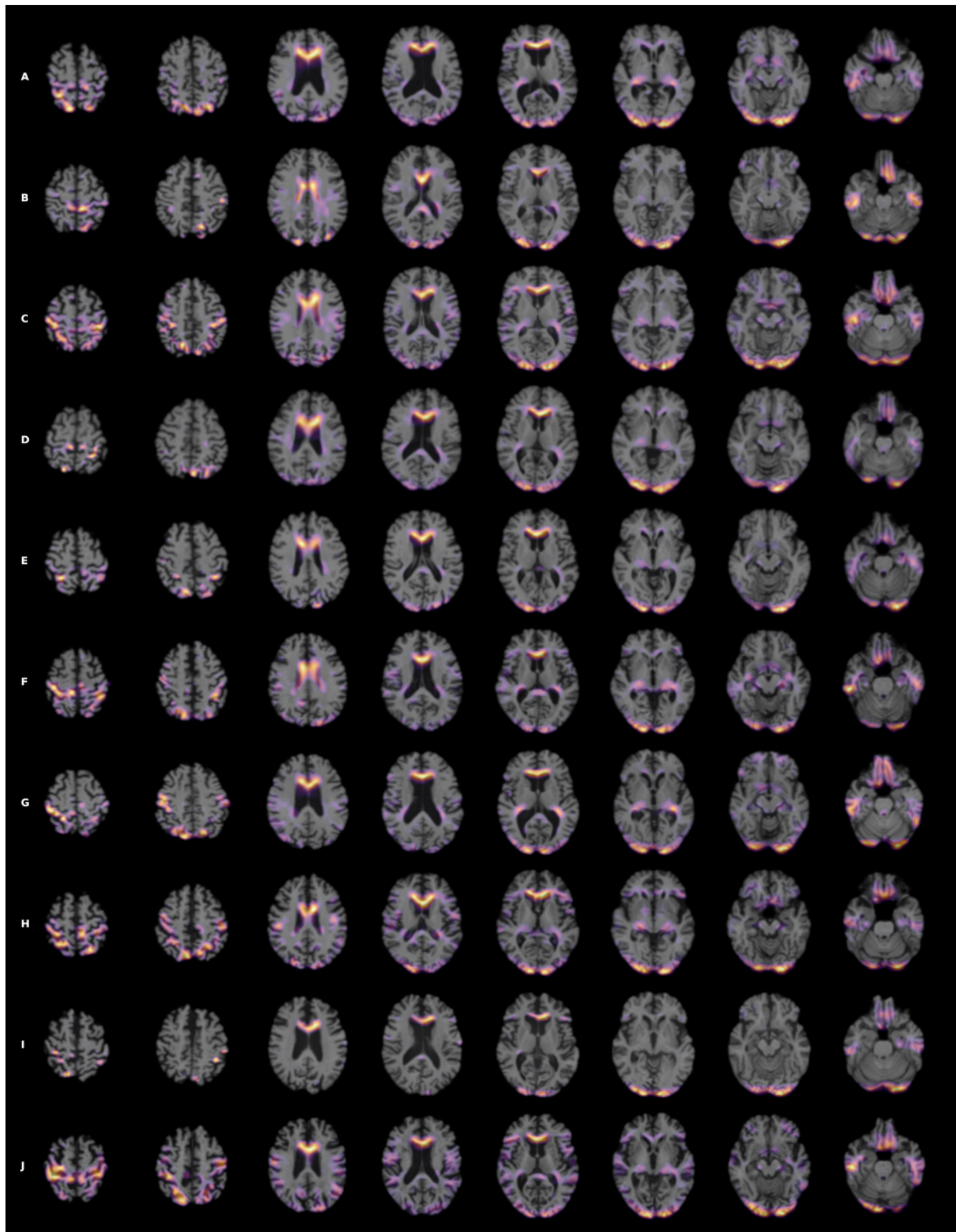



Figure 7

Heatmap visualization for 10 patients.

3D Grad-CAM heatmaps from the wide residual dual channel network used to predict conversion of MCI to AD. Heat maps were superimposed on individual patient's anatomical MRI of 10 patients.A areas in bright yellow-orange (low to high) color corresponding to voxels with the gradient based on 3D Grad-CAM algorithm at the convolutional layer around 20 pixel resolution.

Summary of participant demographics

	N	Age (years)	Gender
NC	320	75.7 ± 5.6 [60.9, 90.8]	160 F / 160 M
AD	237	75.9 ± 7.9 [56.1, 92.0]	108 F / 129 M
sMCI	333	72.7 ± 7.6 [56.1, 89.8]	137 F / 196 M
pMCI	221	75.2 ± 7.2 [55.6, 89.5]	92 F / 129 M

2 NC = normal cognition
3 AD = Alzheimer's disease
4 sMCI = stable mild cognitive impairment
5 pMCI = progressive mild cognitive impairment
6 F = Female
7 M = Male
8 Age values are mean ± standard deviation [range]

AD vs NC classification to generate weights.

Balanced accuracy and area under the receiver operating characteristic curve of the validation and test datasets obtained using single and dual time point networks with sequential, bottleneck residual and wide residual CNN blocks. Dual timepoint networks were twin (equal structure), non-Siamese (separate weights) and merged using concatenation.

	Model Convolution Style	Validation mean ± SD		Test mean ± SD	
		BA	AUC	BA	AUC
Single	Sequential	0.854 ± 0.027	0.918 ± 0.018	**0.860 ± 0.016**	0.922 ± 0.005
	Bottleneck Residual	0.689 ± 0.020	0.774 ± 0.017	0.727 ± 0.051	0.782 ± 0.052
	Wide Residual	0.835 ± 0.025	0.903 ± 0.027	0.840 ± 0.017	0.917 ± 0.006
Dual	Sequential	0.855 ± 0.014	0.938 ± 0.007	0.876 ± 0.010	0.937 ± 0.012
	Bottleneck Residual	0.772 ± 0.046	0.865 ± 0.037	0.800 ± 0.045	0.869 ± 0.043
	Wide Residual	0.856 ± 0.025	0.942 ± 0.012	**0.887 ± 0.009**	0.933 ± 0.003

2 BA = balanced accuracy.
3 AUC = area under the receiver operating characteristic curve
4 Best test average BAs are highlighted in bold for single channel (sequential) and dual channel
5 (wide residual)

Prediction of AD at 3 years.

Results (BA and AUC mean and standard deviation) of prediction using zero-shot and fine-tuning. For single-timepoint networks, resluts are shown using both the baseline and the 1-year MRIs. For zero-shot learning, each of the 4-fold classification trained weights were used as-is against each of the 4 validation fold sets for prediction (16 attempts) and against the prediction test set (4 attempts, with best result also shown). For fine-tuning, the weights from the best test zero-shot result were used as starting weights for training against each of the 4-fold validation sets.

	Model Convolution style	Validation mean ± SD				Test mean ± SD			
		Zero-shot		Fine-tuning		Zero-shot		Fine-tuning	
		BA	AUC	BA	AUC	BA (best)	AUC	BA	AUC
Single (baseline)	Sequential	0.728 ± 0.043	0.790 ± 0.043	0.746 ± 0.0033	0.805 ± 0.028	0.765 ± 0.021 (0.79)	0.831 ± 0.015	0.754 ± 0.026	0.834 ± 0.015
	Wide Residual	0.699 ± 0.034	0.775 ± 0.034	0.700 ± 0.017	0.774 ± 0.022	0.706 ± 0.031 (0.79)	0.816 ± 0.024	0.717 ± 0.030	0.816 ± 0.024
Single (1 year)	Sequential	0.750 ± 0.038	0.807 ± 0.039	0.733 ± 0.050	0.814 ± 0.042	0.774 ± 0.013 (0.79)	0.857 ± 0.012	0.728 ± 0.008	0.836 ± 0.001
	Wide Residual	0.729 ± 0.038	0.799 ± 0.028	0.704 ± 0.037	0.782 ± 0.033	0.743 ± 0.029 (0.77)	0.834 ± 0.020	0.719 ± 0.007	0.803 ± 0.001
Dual	Sequential	0.751 ± 0.027	0.808 ± 0.029	0.712 ± 0.027	0.772 ± 0.039	0.795 ± 0.010 (0.80)	0.874 ± 0.009	0.739 ± 0.012	0.828 ± 0.007
	Wide Residual	0.727 ± 0.038	0.806 ± 0.033	0.719 ± 0.018	0.801 ± 0.032	0.753 ± 0.034 (0.79)	0.842 ± 0.010	0.775 ± 0.003	0.834 ± 0.001

2   AUC = area under the receiver operating characteristic curve

3   BA = balanced accuracy

4   SD = standard deviation