Development and Validation of a Deep Learning–Based Automatic Brain Segmentation and Classification Algorithm for Alzheimer Disease Using 3D T1-Weighted Volumetric Images

C.H. Suh, W.H. Shim, S.J. Kim, J.H. Roh, J.-H. Lee, M.-J. Kim, S. Park, W. Jung, J. Sung, and G.-H. Jahng, for the Alzheimer's Disease Neuroimaging Initiative

ABSTRACT

BACKGROUND AND PURPOSE: Limited evidence has suggested that a deep learning automatic brain segmentation and classification method, based on T1-weighted brain MR images, can predict Alzheimer disease. Our aim was to develop and validate a deep learning–based automatic brain segmentation and classification algorithm for the diagnosis of Alzheimer disease using 3D T1-weighted brain MR images.

MATERIALS AND METHODS: A deep learning–based algorithm was developed using a dataset of T1-weighted brain MR images in consecutive patients with Alzheimer disease and mild cognitive impairment. We developed a 2-step algorithm using a convolutional neural network to perform brain parcellation followed by 3 classifier techniques including XGBoost for disease prediction. All classification experiments were performed using 5-fold cross-validation. The diagnostic performance of the XGBoost method was compared with logistic regression and a linear Support Vector Machine by calculating their areas under the curve for differentiating Alzheimer disease from mild cognitive impairment and mild cognitive impairment from healthy controls.

RESULTS: In a total of 4 datasets, 1099, 212, 711, and 705 eligible patients were included. Compared with the linear Support Vector Machine and logistic regression, XGBoost significantly improved the prediction of Alzheimer disease ($P < .001$). In terms of differentiating Alzheimer disease from mild cognitive impairment, the 3 algorithms resulted in areas under the curve of 0.758–0.825. XGBoost had a sensitivity of 68% and a specificity of 70%. In terms of differentiating mild cognitive impairment from the healthy control group, the 3 algorithms resulted in areas under the curve of 0.668–0.870. XGBoost had a sensitivity of 79% and a specificity of 80%.

CONCLUSIONS: The deep learning–based automatic brain segmentation and classification algorithm allowed an accurate diagnosis of Alzheimer disease using T1-weighted brain MR images. The widespread availability of T1-weighted brain MR imaging suggests that this algorithm is a promising and widely applicable method for predicting Alzheimer disease.

ABBREVIATIONS: AD = Alzheimer disease; ADNI = Alzheimer's Disease Neuroimaging Initiative; AUC = area under the curve; CNN = convolutional neural network; MCI = mild cognitive impairment; OASIS = Open Access Series of Imaging Studies; SVM = Support Vector Machine

Alzheimer disease (AD) is the most common cause of dementia, with mild cognitive impairment (MCI) regarded as a transitional state between normal cognition and early stages of dementia.[1] Although current therapeutic and preventive options are only moderately effective, a reliable decision-making diagnostic approach is important during early stages of AD.[2,3] The guidelines of the National Institute on Aging–Alzheimer's Association suggest that MR imaging is a supportive imaging tool in the diagnostic work-up of patients with AD and MCI.[2,3] Imaging biomarkers play an important role in the diagnosis of AD, both in

Received March 4, 2020; accepted after revision August 7.

From the Department of Radiology and Research Institute of Radiology (C.H.S., W.H.S., S.J.K.), Department of Neurology (J.H.R., J.-H.L.), and Health Screening and Promotion Center (M.-J.K.), Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea; Department of Physiology (J.H.R.), Korea University College of Medicine, Seoul, Republic of Korea; VUNO Inc (S.P., W.J., J.S.), Seoul, Republic of Korea; and Department of Radiology (G.-H.J.), Kyung Hee University Hospital at Gangdong, College of Medicine, Kyung Hee University, Seoul, Republic of Korea.

C.H. Suh and W.H. Shim contributed equally to this article.

This work was supported by a grant from the Institute for Information and Communications Technology Promotion, funded by the Korean government (C0510-18-1001; Intelligent SW Technology Development for Medical Data Analysis); a grant from the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (HI11C1238 or A111282); and a grant from the Basic Science Research Program through the National Research Foundation of Korea grant funded by the Korea government (2014R1A2A2A01002728).

Please address correspondence to Sang Joon Kim, MD, PhD, Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-Gu, Seoul 05505, Republic of Korea; e-mail: sjkimjb5@gmail.com

Indicates open access to non-subscribers at www.ajnr.org
Indicates article with supplemental online appendix and table.
http://dx.doi.org/10.3174/ajnr.A6848

Table 1: Characteristics of the development and datasets[a]

	Asan Medical Center	Kyung Hee University Hospital in Gangdong	ADNI	OASIS
No. of patients	1099	212	711	705
Age (mean) (yr)	65 ± 13	70 ± 9	76 ± 7	68 ± 10
No. of male patients	500 (45)	52 (25)	412 (58)	403 (57)
No. of female patients	599 (55)	160 (75)	299 (42)	302 (43)
Classification				
AD	161 (15)	68 (32)	178 (25)	145 (21)
Education (yr)	9.9 (4.8)	NA	14.5 (3.4)	14.0 (3.2)
MMSE score	18.5 (4.7)	17.4 (5.3)	22.8 (3.1)	24.4 (5.1)
Clinical Dementia Rating	1.00 (0.49)	1.10 (0.47)	0.73 (0.34)	0.68 (0.28)
Global Deterioration Scale	NA	NA	1.7 (1.4)	3.2 (7.3)
MCI	363 (33)	63 (30)	317 (45)	0
Education (yr)	10.1 (5.0)	NA	15.9 (2.5)	
MMSE score	24.9 (3.6)	25.7 (3.7)	26.4 (2.1)	
Clinical Dementia Rating	0.51 (0.09)	0.61 (1.16)	0.5	
Global Deterioration Scale	NA	NA	1.5 (1.3)	
Healthy control	575 (52)	81 (38)	216 (30)	560 (79)
Education (yr)	NA	NA	16.2 (2.8)	15.2 (2.7)
MMSE score	29.5 (0.5)	27.7 (2.5)	29.1 (1.0)	28.8 (3.2)
Clinical Dementia Rating	NA	0.24 (0.26)	0	0
Global Deterioration Scale	NA	NA	0.8 (1.1)	1.3 (4.0)

Note:—MMSE indicates Mini-Mental State Examination; NA, not available.
[a] Unless otherwise indicated, data are reported as number (%).

the research field and in clinical practice. The identification of amyloid and the τ PET ligand provided huge advances in understanding the pathophysiologic mechanisms underlying AD and its early diagnosis, even in the preclinical or prodromal stage.[4-6] Although amyloid and τ PET are more sensitive and specific for the diagnosis of AD, they are expensive to perform, have limited availability, and require ionizing radiation, limiting their use in clinical practice. CSF amyloid and τ are also important biomarkers that could be used for AD diagnostics in the clinical research setting.[3,7-9] However, CSF AD biomarkers also have limited availability. MR imaging, however, is widely available and used in standard practice to support the diagnosis of AD and to exclude other causes of cognitive impairment, including stroke, vascular dementia, normal-pressure hydrocephalus, and inflammatory and neoplastic conditions.

3D T1-weighted volumetric MR imaging is the most important MR imaging tool in the diagnosis of AD. 3D volumetry has long been used as a morphologic diagnostic tool for AD, not only as a visual assessment or manual segmentation but for semiautomatic and automatic segmentation. Examples include semiautomatic structural changes on MR imaging,[10] automated hippocampal volumetry,[11] entorhinal cortex atrophy,[12] and changes in pineal gland volume.[13] Although user-friendly automated segmentation algorithms were first introduced 20 years ago, evidence supporting the use of 3D volumetry in clinical practice is currently insufficient. Visual assessment requires experience, and automatic 3D volumetry requires a long acquisition time.

To our knowledge, limited evidence has suggested that a deep learning automatic brain segmentation and classification method, based on T1-weighted brain MR images, can predict AD.[14] Currently available algorithms have low clinical feasibility because of the long processing time for brain segmentation, and the classification algorithm based on T1-weighted brain MR images needs to be validated in a large external dataset. The purpose of this study was to develop and validate a deep learning–based automatic brain segmentation and classification algorithm for the diagnosis of AD using 3D T1-weighted brain MR images.

MATERIALS AND METHODS

This study was approved by the institutional review boards of all participating institutions, which waived the requirement for informed consent due to the retrospective design of this study.

Development and Validation Dataset

The deep learning–based automatic brain segmentation and classification algorithm was developed using a dataset of T1-weighted brain MR images from consecutive patients with AD and MCI who met the diagnostic criteria. This dataset was derived from consecutive patients who were referred to a neurology memory clinic and underwent brain MR imaging at Asan Medical Center between December 2014 and March 2017. Patients were considered eligible if their electronic medical records were available, and they had no treatment history of antidementia or psychoactive drugs and no history of neurologic or psychiatric disorders other than AD or MCI. Clinical diagnosis served as the reference standard for AD and MCI, which were diagnosed in all patients by 2 experienced neurologists on the basis of the diagnostic guidelines of the National Institute on Aging–Alzheimer's Association workgroups.[3,7] During the same period, healthy controls were enrolled at Asan Medical Center Health Screening and Promotion Center. Healthy controls were recruited with the following inclusion criteria: no memory impairment, no history of neurologic or psychiatric disorders, and no history of being treated with antidementia or psychoactive drugs.

The patients and healthy controls from Kyung Hee University Hospital in Gangdong who met the same eligibility criteria were evaluated. To externally validate the algorithm using public datasets, we used the Alzheimer Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS) datasets. Their final diagnoses were downloaded from the ADNI web portal (adni.loni.ucla.edu)[15] and the OASIS web portal (oasis-brains.org), respectively. Patients included in these datasets met the same eligibility criteria. None of the brain MR images in datasets overlapped the images in the other datasets. The characteristics of the datasets are shown in Table 1.

All classification experiments were performed using 5-fold cross-validation. Each of the 4 datasets was divided into 5 folds. For each fold containing 4/5 and 1/5 of the training and validation split, respectively, the training set was further partitioned evenly into 5 segments to obtain an ensemble of 5 models, which was then evaluated on the remaining 1/5 validation data. Areas under the curve (AUCs), sensitivity, specificity, positive predictive value, and negative predictive value were used as the evaluation metrics.

MR Imaging Protocol

The MR imaging data in this study were obtained using various MR imaging machines at multiple institutions. MRIs at Asan Medical Center were performed on 3T units (Ingenia; Philips Healthcare) using a 32-channel sensitivity encoding head coil. High-resolution anatomic 3D volume images were obtained in the sagittal plane using a 3D gradient-echo T1-weighted sequence. The detailed parameters included a TR of 9.6 ms, TE of 4.6 ms, a flip angle of 8°, an FOV of 224 × 224 mm, section thickness of 1 mm with no gap, and a matrix size of 224 × 224.

MRIs at Kyung Hee University Hospital in Gangdong were performed on a 3T MR imaging scanner (Achieva; Philips Healthcare) using a dedicated 8-element phased array sensitivity encoding head coil. 3D T1-weighted sagittal images were acquired using an MPRAGE sequence with imaging parameters that included a TR of 9.9 ms, a TE of 4.6 ms, a flip angle of 8°, an FOV of 240 × 240 mm, a section thickness of 1 mm, a matrix size of 240 × 240, and a resolution of 1.00 × 1.00 × 1.00 mm^3. In the ADNI dataset, the section thickness was 1.2 mm with no gaps. In the OASIS dataset, the section thickness was 1.25 mm with no gaps.

Development of Deep Learning–Based Automatic Classification Algorithm

Brain Parcellation Module. The proposed deep learning–based AD classification system consisted of a deep convolutional neural network (CNN) module and an XGBoost module (https://hackernoon.com/want-a-complete-guide-for-xgboost-model-in-python-using-scikit-learn-sc11f31bq). The deep CNN module parcellated each brain into 82 areas. The proposed deep CNN had a 2.5 channel HighResnet architecture (https://github.com/NifTK/NiftyNet/tree/dev/demos/brain_parcellation), consisting of 44 convolution layers without a strided convolution or pooling layer. The HighResnet architecture, in which layers were stacked as deep as possible using atrous convolution rather than pooling or stride, has been shown to perform brain parcellation well.[16] 2.5D CNN is a method designed to use 3D information while still using 2D CNN architecture. This method concatenates a target section and other slices around the target in the channel dimension and uses it as the input to the network. This method is widely used for medical images that include 3D imaging data.[17]

HighResnet is a network of deeply stacked blocks with residual connections. The residual connection is a method proposed to solve the degradation problem, in which accuracy is saturated as the depth of the network increases.[18] The residual connection helps in effective training, even if the layer blocks are deeply stacked. A neural network with n residual connections was found to have 2^n unique pathways.[19] Thus, a network with residual connections has the same effect as using receptive fields of various sizes without having a fixed receptive field.[20] Details of the brain parcellation module are described in the On-line Appendix.

Volume-Based AD Classification Algorithm. On the basis of this parcellated brain volume information, the XGBoost module classified patients into the AD, MCI, and healthy control groups. We compared our AD classification method against logistic regression and the linear Support Vector Machine (SVM). Both methods have been implemented by Scikit-learn (Version 0.21.0; https://scikit-learn.org/). The detailed network structure of the parcellation CNN module is summarized in Fig 1. The 2 modules were cascaded for use as a fully automated classification system. The network was trained using an ADAM optimizer[21] with an initial learning rate of 0.001. The exponential decay rates for the first- and second-moment estimates were 0.9 and 0.999, respectively.

Both modules were coded in Python (Version 2.7; Python Software Foundation). The parcellation CNN module was implemented using Tensorflow libraries (Version 1.12; https://www.tensorflow.org), whereas the classification XGBoost module was implemented using DMLC XGBoost packages (Version 0.80; https://xgboost.ai/ and https://github.com/dmlc/xgboost/blob/master/CITATION).

This study did not use an end-to-end approach to classify AD. Rather, the entire system was divided into a parcellation module using a CNN and a classification module using XGBoost. The features extracted from the CNN activation maps are difficult to interpret medically, whereas the volumes of brain regions are directly associated with the degree of cortical atrophy due to AD. In addition, differently distributed volumes can distinguish among AD, MCI, and healthy controls. In neurodegeneration research, normalization of regional volume by intracranial volume is crucial to reduce interindividual variation. To measure whole-brain volumes, we developed a brain-extraction method, which is another deep learning–based semantic segmentation algorithm. We divided raw volumes of brain parcellation by the whole-brain volume. In addition, to remove the age-related effects of brain volumes and reflect sex matching, we composed 82 volumes, age, and sex (0 or 1) as input data for classification. It is a multivariate approach of age and sex matching. Transformation of each T1-weighted brain MR image into the volume of each brain region reduces the dimensionality of the data. When the data dimensionality is relatively small, a classifier using the boosting technique is efficient.[22] Therefore, AD, MCI, and healthy controls were classified using XGBoost.

Boosting is a method by which weak classifiers can be grouped into sets, with these ensembles used to predict results. XGBoost is a tree-boosting algorithm, using an ensemble model called a classification and a regression tree to create a tree classifier. Tree boosting is a highly effective and widely used machine learning method. The hyperparameters for XGBoost learning were set at a maximum depth of 5, 102 estimators, and a learning rate of 0.9.

Evaluation of Algorithms and Statistical Analyses

On the basis of T1-weighted brain MR images, the trained deep learning–based automatic classification algorithm generated continuous probabilities, ranging from 0 to 1, that patients had AD. The primary outcome was to investigate the diagnostic performance of

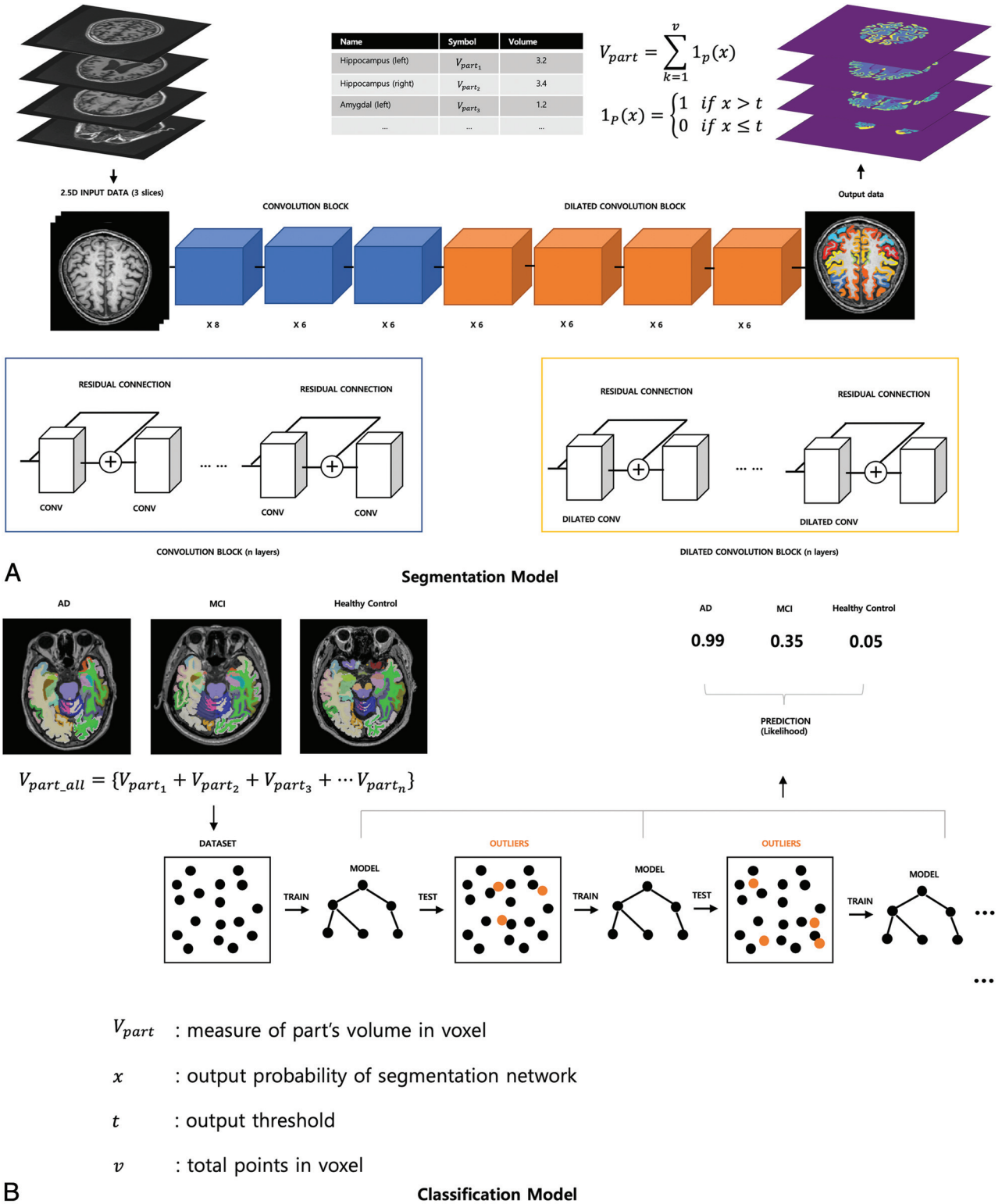

FIG 1. Network architecture of the brain parcellation and classification model. CONV indicates convolution.

the algorithm in differentiating AD from MCI and MCI from healthy controls. The secondary outcome was to investigate the diagnostic performance of the algorithm in differentiating AD from healthy controls. The impact of each feature (volume of each brain region) on the AD prediction model was reported using Shapley values (Fig 2), in which the impact of a feature is defined as the change in the expected output of the model when a feature is, compared with when it is not, observed.[23]

The XGBoost method was compared with 2 other commonly used classification methods for the prediction of AD, logistic regression and linear SVM. The diagnostic performance of the 3 methods was compared using the method of Delong et al[24] to

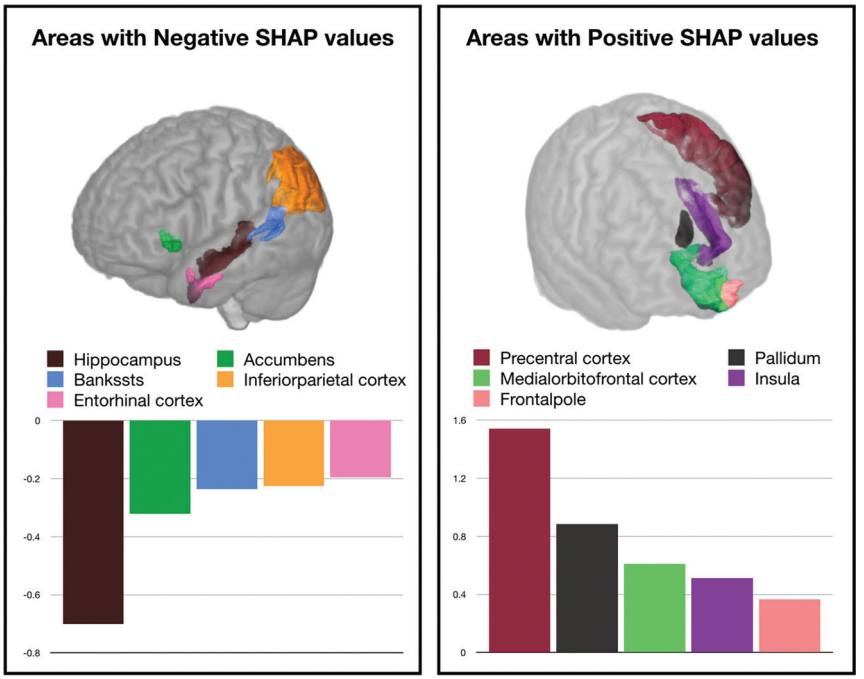

FIG 2. The impact of feature (volume of each brain region) on the AD prediction model, as represented by Shapley values, in which the impact of a feature is defined as the change in the expected output of the model when a feature is observed versus unknown. *A*, Visualization of the top 5 brain regions representing feature impacts pushing the decision of the model to AD, along with average feature impact. *B*, Visualization of the top 5 brain regions representing feature impacts pushing the decision of the model to healthy controls, along with average feature impact. Bankssts indicates banks of the superior temporal sulcus; SHAP, Shapley Additive Explanations (https://pbiecek.github.io/ema/shapley.html).

calculate the standard error of the AUC and the difference among the 3 AUCs. Optimal cutoff probabilities for differentiating AD or MCI were obtained from receiver operating characteristic curves, with the sensitivity, specificity, positive predictive value, negative predictive value, and AUC calculated using the Youden index,[25] defined as sensitivity + specificity −1, with values ranging from −1 to +1. The parcellation module was evaluated with a mean Dice Similarity Coefficient using the ground truth segmentation mask of FreeSurfer (http://surfer.nmr.mgh.harvard.edu). All statistical analyses were performed using MedCalc, Version 18.6 (MedCalc Software), with $P < .05$ defined as statistically significant.

RESULTS

Patient Demographics

Of the 1099 eligible patients who underwent T1-weighted MR imaging at the Asan Medical Center, 161 were diagnosed with probable AD, 363 were diagnosed with MCI, and 575 were classified as healthy controls (Table 1). The mean ages of these 3 groups were 75 ± 8 years, 69 ± 10 years, and 57 ± 9 years, respectively, and there was a statistically difference ($P < .01$). In 212 patients from the dataset of Kyung Hee University Hospital in Gangdong, 68 patients were diagnosed with probable AD, 63 were diagnosed with MCI, and 81 were classified as healthy controls. The mean ages of these 3 groups were 75 ± 8 years, 70 ± 8 years, and 65 ± 9 years, respectively. and there was a statistically significant difference ($P < .01$). The ADNI dataset included 178 patients diagnosed with AD, 317 diagnosed with MCI, and 216 healthy controls; their mean ages were 76 ± 8 years, 75 ± 8 years, and 77 ± 5 years, respectively. The OASIS dataset included 145 patients diagnosed with AD and 560 healthy controls; their mean ages were 74 ± 8 years and 70 ± 9 years, respectively.

Diagnostic Performance in AD versus MCI

In the Asan Medical Center dataset, the AUCs for logistic regression, linear SVM, and XGBoost were 0.770 (95% CI, 0.761–0.779), 0.772 (95% CI, 0.761–0.782), and 0.803 (95% CI, 0.802–0.805), respectively (Table 2). Use of XGBoost significantly improved the prediction of AD compared with the linear SVM ($P < .001$) and logistic regression ($P < .001$). Because XGBoost showed the highest AUC, this method was chosen to provide the optimal cutoff value. XGBoost had a sensitivity of 71% (95% CI, 69%–72%) and a specificity of 74% (95% CI, 74%–74%), with an optimal cutoff value of 0.613 (On-line Table).

In the dataset of the Kyung Hee University Hospital in Gangdong, the AUCs for logistic regression, linear SVM, and XGBoost were 0.798 (95% CI, 0.775–0.822), 0.804 (95% CI, 0.783–0.824), and 0.825 (95% CI, 0.810–0.840). In the ADNI dataset, the AUCs for logistic regression, linear SVM, and XGBoost were 0.706 (95% CI, 0.702–0.710), 0.700 (95% CI, 0.695–0.704), and 0.758 (95% CI, 0.755–0.760), respectively.

Diagnostic Performance in MCI versus Healthy Controls

In the Asan Medical Center dataset, the AUCs for logistic regression, linear SVM, and XGBoost were 0.812 (95% CI, 0.806–0.817), 0.830 (95% CI, 0.821–0.840), and 0.870 (95% CI, 0.868–0.872), respectively (Table 2). Use of XGBoost significantly improved the prediction of AD compared with the linear SVM ($P < .001$) and logistic regression ($P < .001$). XGBoost had a sensitivity of 79% (95% CI, 78%–79%) and a specificity of 80% (95% CI, 79%–81%), with an optimal cutoff value of 0.016 (On-line Table).

In the dataset of the Kyung Hee University Hospital in Gangdong, the AUCs for logistic regression, linear SVM, and XGBoost were 0.692 (95% CI, 0.678–0.706), 0.687 (95% CI, 0.669–0.706), and 0.705 (95% CI, 0.699–0.712), respectively. In the ADNI dataset, the AUCs for logistic regression, linear SVM, and XGBoost were 0.698 (95% CI, 0.686–0.710), 0.702 (95% CI, 0.697–0.708), and 0.668 (95% CI, 0.664–0.671), respectively. Th diagnostic performance of the algorithm in differentiating AD from healthy controls is shown in Table 2 and the On-line Table.

Table 2: Diagnostic performance of logistic regression, the linear Support Vector Machine, and the deep learning–based automatic classification algorithm in the datasets[a]

	Logistic Regression	Linear SVM	XGBoost	*P* Value[b]	*P* Value[c]
AD vs MCI					
Asan Medical Center	0.770 (0.761–0.779)	0.772 (0.761–0.782)	0.803 (0.802–0.805)	<.001	<.001
Kyung Hee University Hospital at Gangdong	0.798 (0.775–0.822)	0.804 (0.783–0.824)	0.825 (0.810–0.840)	.018	.030
ADNI	0.706 (0.702–0.710)	0.700 (0.695–0.704)	0.758 (0.755–0.760)	<.001	<.001
MCI vs healthy control					
Asan Medical Center	0.812 (0.806–0.817)	0.830 (0.821–0.840)	0.870 (0.868–0.872)	<.001	<.001
Kyung Hee University Hospital at Gangdong	0.692 (0.678–0.706)	0.687 (0.669–0.706)	0.705 (0.699–0.712)	.029	.023
ADNI	0.698 (0.686–0.710)	0.702 (0.697–0.708)	0.668 (0.664–0.671)	<.001	<.001
AD vs healthy controls					
Asan Medical Center	0.953 (0.949–0.958)	0.960 (0.958–0.963)	0.982 (0.980–0.985)	<.001	<.001
Kyung Hee University Hospital at Gangdong	0.905 (0.889–0.921)	0.911 (0.903–0.920)	0.940 (0.933–0.947)	<.001	<.001
ADNI	0.863 (0.856–0.870)	0.860 (0.857–0.863)	0.885 (0.879–0.891)	<.001	<.001
OASIS[d]	0.826 (0.817–0.835)	0.820 (0.809–0.832)	0.840 (0.837–0.844)	.001	<.001

[a] Data are AUC (95% CI).
[b] *P* values: between logistic regression and XGBoost.
[c] *P* values: between linear SVM and XGBosst.
[d] OASIS dataset included only AD and healthy controls.

Performance Evaluation of Brain Parcellation Module

Dice Similarity Coefficients for Asan Medical Center, ADNI, and OASIS datasets were 82.0 (95% CI, 81.6–82.4), 82.3 (95% CI, 81.5–83.1), and 82.0 (95% CI, 81.6–82.4), respectively. This performance is almost identical to the 82.05, on average, reported by Li et al.[16]

DISCUSSION

The present study describes the development and validation of a deep learning–based automatic brain segmentation and classification algorithm using T1-weighted brain MR images for the diagnosis of AD. This algorithm resulted in the accurate diagnosis of AD, with AUCs of 0.758–0.825 in differentiating AD from MCI and AUCs of 0.668–0.870 in differentiating MCI from healthy controls. Because of the widespread availability of T1-weighted brain MR imaging, the deep learning–based automatic brain segmentation and classification algorithm is a promising and widely applicable method for prediction of AD.

The CNN parcellation module developed in this study successfully mimicked FreeSurfer,[26] with only 20 seconds required for parcellation and classification of each MR image. One of the disadvantages of previous methods for parcellation, including FreeSurfer and NeuroQuant (CorTechs Labs), was their long processing times (FreeSurfer, 7 hours; NeuroQuant, 5–7 minutes).[27,28] In addition, our deep learning–based automatic brain segmentation and classification algorithm (XGBoost) was robust across various clinical settings, even in public datasets, showing improved diagnostic performance for the prediction of AD compared with the linear SVM and logistic regression. XGBoost can easily handle sparse data using sparsity-aware algorithms and is scalable to various tasks[22] in medicine, including AD classification, medical text data, and temporal data. The gradient-boosting algorithm constructed the new base learners to be maximally correlated with the negative gradient of the loss function, which is associated with many decision trees (weak learners). The gradient boosting algorithm consistently provided greater accuracy than conventional single, strong, machine learning models. Because the Dice Similarity Coefficients for the Asan Medical Center, ADNI, and OASIS datasets were similar, XGBoost may contribute to performance differences among XGBoost, SVM, and logistic regression.

The algorithm we developed was based on brain volumes determined on T1-weighted brain MR images. This algorithm yielded probabilities of 0–1 for each patient. The optimal cutoff value was 0.613, showing a sensitivity of 71% (95% CI, 69%–72%) and a specificity of 74% (95% CI, 74%–74%) in predicting AD differentiation from MCI. In clinical practice, it is difficult to predict AD using MR imaging, though several imaging findings may be predictive of advanced AD. MR imaging findings during the early stages of AD are subtle, with visual assessments of these findings being subjective. Use of our high-speed, accurate deep learning–based automatic brain segmentation and classification algorithm could predict the likelihood of AD in patients with cognitive impairment or when screening individuals in daily clinical practice. Moreover, the present study demonstrated high diagnostic performance of the algorithm in differentiating AD from healthy controls (AUC = 0.840–0.982) and MCI from healthy controls (AUC = 0.668–0.870). Thus, our results may broaden the clinical utility of a deep learning–based automatic brain segmentation and classification algorithm for patients with memory impairment.

Among the various imaging methods available for evaluating AD, T1-weighted brain MR imaging and FDG-PET MR imaging have been widely validated and have shown clinical efficacy.[14,29] For example, an ensemble learning system for classification of AD, MCI, and healthy controls was developed using an ADNI dataset, and a parameter-efficient deep learning approach was found to be highly accurate (AUC = 0.925) in predicting conversion from MCI to AD in an ADNI dataset.[14] Similarly, the accuracy of a deep learning algorithm for early prediction of AD using ^{18}F. FDG-PET results was found to be 0.98 (95% CI, 0.94–1.00).[29] These studies mainly focused on predicting the early conversion to AD among patients with MCI. By contrast, our study demonstrated that our algorithm was accurate in differentiating AD from MCI (AUCs = 0.758–0.825) and MCI from healthy controls (AUC = 0.668–0.870), which may be due to the large overlap between AD and MCI. However, our findings were validated externally in large patient cohorts. In the Kaggle 2016 competition (a machine learning neuroimaging

challenge for automated diagnosis of mild cognitive impairment), the winner of the competition attempted to quantify the prediction accuracy of multiple morphologic MR imaging features and achieved a precision of 76% for the class AD and a precision of 45%–64% for the class MCI, which was lower than our results.[30]

This study externally validated our deep learning–based automatic brain segmentation and classification algorithm using 3 different test datasets. A recent analysis reported that only 31 of 516 (6%) studies included external validation.[31] Evaluation of the clinical performance of a diagnostic or predictive artificial intelligence model requires the analysis of external data from a clinical cohort that appropriately represents the target patient population, avoiding overestimation of the initial results because of overfitting and spectrum bias.[32] Our results showed that the diagnostic performance of our algorithm was similar for internal (AUC = 0.803) and external (AUC = 0.758–0.825) datasets, indicating no overfitting.

The present study had several limitations. First, this study was based on retrospective data from selected patient groups and did not include patients with non-AD neurodegenerative diseases. This study, however, was not intended to develop an all-inclusive tool to differentiate various causes of cognitive impairment, suggesting that application of this algorithm to such populations may be limited. Further validation with larger, prospectively collected test datasets may be necessary to determine whether our algorithm is applicable to various types of cognitive impairment.[33] Second, the diagnostic criteria of AD were based on the clinical diagnosis; therefore, these might be different from a diagnosis based on amyloid PET or τ PET. However, our study was based on retrospective data from the clinical field; thus, we could use only diagnostic criteria of AD based on clinical diagnosis. Third, further research is required to assess the clinical benefits of the deep learning–based automatic classification algorithm in predicting the prognosis and helping to manage patients with AD. In addition, longitudinal outcome studies evaluating the likelihood of decline or progression to MCI or AD in an individual using longitudinal MR imaging data are necessary.

CONCLUSIONS

The deep learning–based automatic brain segmentation and classification algorithm developed in this study was accurate in diagnosing AD using T1-weighted brain MR images. The widespread availability of T1-weighted brain MR imaging indicates that this algorithm may be a promising and widely applicable method for prediction of AD.

Disclosures: Sang Joon Kim—*RELATED*: *Grant*: funded by the Korean government (C0510-18-1001). Jee Hoon Roh—*RELATED*: *Grant*: from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute funded by the Ministry of Health and Welfare, Republic of Korea (HI14C3319).* Sejin Park—*UNRELATED*: *Employment:* VUNO Inc. Jinkyeong Sung—*UNRELATED*: *Employment*: VUNO Inc. Wonmo Jung—*UNRELATED*.* Money paid to the institution.

REFERENCES

1. Petersen RC, Negash S. **Mild cognitive impairment: an overview.** *CNS Spectr* 2008;13:45–53 CrossRef Medline
2. Dubois B, Feldman HH, Jacova C, et al. **Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria.** *Lancet Neurol* 2007;6:734–46 CrossRef Medline
3. McKhann GM, Knopman DS, Chertkow H, 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;7:263–69 CrossRef Medline
4. Jack CR Jr, Bennett DA, Blennow K, et al; Contributors. **NIA-AA research framework: toward a biological definition of Alzheimer's disease.** *Alzheimers Dement* 2018;14:535–62 CrossRef Medline
5. Johnson KA, Schultz A, Betensky RA, et al. **Tau positron emission tomographic imaging in aging and early Alzheimer disease.** *Ann Neurol* 2016;79:110–19 CrossRef Medline
6. Jack CR Jr, Bennett DA, Blennow K, et al. **A/T/N: an unbiased descriptive classification scheme for Alzheimer disease biomarkers.** *Neurology* 2016;87:539–47 CrossRef Medline
7. Albert MS, DeKosky ST, Dickson D, 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;7:270–79 CrossRef Medline
8. Sperling RA, Aisen PS, Beckett LA, 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 CrossRef Medline
9. Jack CR Jr, Albert MS, Knopman DS, et al. **Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease.** *Alzheimers Dement* 2011;7:257–62 CrossRef Medline
10. McEvoy LK, Fennema-Notestine C, Roddey JC, et al; Alzheimer's Disease Neuroimaging Initiative. **Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment.** *Radiology* 2009;251:195–205 CrossRef Medline
11. Colliot O, Chetelat G, Chupin M, et al. **Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging using automated segmentation of the hippocampus.** *Radiology* 2008;248:194–201 CrossRef Medline
12. Enkirch SJ, Traschutz A, Muller A, et al. **The ERICA score: an MR imaging-based visual scoring system for the assessment of entorhinal cortex atrophy in Alzheimer disease.** *Radiology* 2018;288:226–333 CrossRef Medline
13. Matsuoka T, Imai A, Fujimoto H, et al. **Reduced pineal volume in Alzheimer disease: a retrospective cross-sectional MR imaging study.** *Radiology* 2018;286:239–48 CrossRef Medline
14. Spasov S, Passamonti L, Duggento A, et al; Alzheimer's Disease Neuroimaging Initiative. **A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease.** *Neuroimage* 2019;189:276–87 CrossRef Medline
15. Mueller SG, Weiner MW, Thal LJ, et al. **The Alzheimer's Disease Neuroimaging Initiative.** *Neuroimaging Clin N Am* 2005;15:869–77. xi-xii CrossRef Medline
16. Li W, Wang G, Fidon L, et al. **On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task.** In: *Proceedings of the International Conference on Information Processing in Medical Imaging*, Boone, North Carolina; June 25–30, 2017
17. Roth HR, Lu L, Seff A, et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. In: *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*: Springer; 2014;520–27
18. He K, Zhang X, Ren S, et al. **Deep residual learning for image recognition.** In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada; June 27–30, 2016 CrossRef

19. Veit A, Wilber M, Belongie S. **Residual networks are exponential ensembles of relatively shallow networks.** October 2016. *arXi.org*. https://arxiv.org/abs/1605.06431v1. Accessed Aug 1, 2019
20. Luo W, Li Y, Urtasun R, et al. **Understanding the effective receptive field in deep convolutional neural networks.** January 2017. *arXiv.org*. https://arxiv.org/abs/1701.04128. Accessed Aug 1, 2019
21. Kingma DP, Ba J. **Adam: a method for stochastic optimization.** December 2014. *arXiv.org* https://arxiv.org/abs/1412.6980. Accessed Aug 1, 2019
22. Chen T, Guestrin C. **XGBoost: a scalable tree boosting system.** In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California. August 2016; 785–94
23. Strumbelj E, Kononeko I. **An efficient explanation of individual classifications using game theory.** *Journal of Machine Learning Research* 2010;11:1–18 CrossRef
24. DeLong ER, DeLong DM, Clarke-Pearson DL. **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.** *Biometrics* 1988;44:837–45 Medline
25. Youden WJ. **Index for rating diagnostic tests.** *Cancer* 1950;3:32–35 CrossRef Medline
26. Dale AM, Fischl B, Sereno MI. **Cortical surface-based analysis, I: segmentation and surface reconstruction.** *Neuroimage* 1999;9:179–94 CrossRef Medline
27. Ochs AL, Ross DE, Zannoni MD, et al; Alzheimer's Disease Neuroimaging Initiative, **Comparison of automated brain volume measures obtained with NeuroQuant and FreeSurfer.** *J Neuroimaging* 2015;25:721–27CrossRef Medline
28. Persson K, Barca ML, Cavallin L, et al. **Comparison of automated volumetry of the hippocampus using NeuroQuant and visual assessment of the medial temporal lobe in Alzheimer's disease.** *Acta Radiol* 2018;59:997–1001 CrossRef Medline
29. Ding Y, Sohn JH, Kawczynski MG, et al. **A deep learning model to predict a diagnosis of Alzheimer disease using (18)F-FDG PET of the brain.** *Radiology* 2019;290:456–64 CrossRef Medline
30. Dimitriadis SI, Liparas D, Tsolanki MN, et al. **Random forest feature selection, fusion and ensemble strategy: combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: from the alzheimer's disease neuroimaging initiative (ADNI) database.** *Journal of Neuroscience Methods* 2018;302:14–23
31. Kim DW, Jang HY, Kim KW, et al. **Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers.** *Korean J Radiol* 2019;20:405–10 CrossRef Medline
32. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
33. Park SH. **Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance.** *Radiology* 2019;290:272–73 CrossRef Medline