ARTICLE OPEN

Check for updates

A high-generalizability machine learning framework for predicting the progression of Alzheimer's disease using limited data

Caihua Wang[1]✉, Yuanzhong Li[1]✉, Yukihiro Tsuboshita[1], Takuya Sakurai[1], Tsubasa Goto[1], Hiroyuki Yamaguchi[2,3], Yuichi Yamashita[2], Atsushi Sekiguchi[4], Hisateru Tachimori[5,6] and for the Alzheimer's Disease Neuroimaging Initiative*

Alzheimer's disease is a neurodegenerative disease that imposes a substantial financial burden on society. A number of machine learning studies have been conducted to predict the speed of its progression, which varies widely among different individuals, for recruiting fast progressors in future clinical trials. However, because the data in this field are very limited, two problems have yet to be solved: the first is that models built on limited data tend to induce overfitting and have low generalizability, and the second is that no cross-cohort evaluations have been done. Here, to suppress the overfitting caused by limited data, we propose a hybrid machine learning framework consisting of multiple convolutional neural networks that automatically extract image features from the point of view of brain segments, which are relevant to cognitive decline according to clinical findings, and a linear support vector classifier that uses extracted image features together with non-image information to make robust final predictions. The experimental results indicate that our model achieves superior performance (accuracy: 0.88, area under the curve [AUC]: 0.95) compared with other state-of-the-art methods. Moreover, our framework demonstrates high generalizability as a result of evaluations using a completely different cohort dataset (accuracy: 0.84, AUC: 0.91) collected from a different population than that used for training.

npj Digital Medicine (2022)5:43 ; https://doi.org/10.1038/s41746-022-00577-x

INTRODUCTION

Alzheimer's disease (AD) is a progressive disorder that causes brain cells to degenerate, and its symptoms, such as memory impairment, greatly impact the activities of daily living of affected patients. As of 2019, there were an estimated 5.8 million patients with AD, and this number is expected to grow to 13.8 million by 2050 in the USA alone[1]. On the other hand, AD has few available treatments, and there has been a high rate of failure in AD drug development programs[2]. New drug development remains challenging, and a large number of such drugs aim to slow the progress of AD at an early stage[3]. Participants with mild cognitive impairment (MCI) who are considered to be in the prodromal stage of AD are generally involved in clinical trials. For such trials, it is important to recruit participants with MCI who are likely to progress to AD, because only 20% of patients with MCI are subsequently diagnosed with AD within 1.5–2 years, while the other 80% remain unchanged or even revert to normal[4,5]. Therefore, an algorithm that can classify stable MCI (sMCI) and progressive MCI (pMCI) is needed to indicate that a patient with MCI will progress to AD within a certain period.

Various machine learning models for predicting the progress of AD have been proposed, and the most frequently utilized dataset, which includes brain images and cognitive test scores, is from the North American Alzheimer's Disease Neuroimaging Initiative[6] (NA-ADNI). Among these models, some[7–10] use only brain images as input; the others[11–17] use multimodal information including not only images but also cognitive scores. A summary of related works is provided in Supplementary Table 1, which shows that while some image-based methods have adopted an end-to-end approach, that is, using voxel-level images as inputs, almost all methods based on both image and clinical information use image features, such as hippocampus volumes, extracted by their own or existing open-source tools. Because most of these models are implemented using deep learning, which requires training of a considerable number of parameters, a large number of training samples is desirable to achieve good generalization performance[18]. Even though the classification accuracies of sMCI and pMCI have reached 75% to 86%, two problems have yet to be solved. The first problem is that the limited data in the NA-ADNI tend to induce overfitting. Compared with ChestX-ray[19] and ImageNet[20], which include samples from tens of thousands to millions of patients and objects, the NA-ADNI contains only about 1000 patients with MCI. The second problem is that, because of the participants' biases, the actual ability of an artificial intelligence (AI) model trained and evaluated based on the NA-ADNI needs to be tested on a completely different dataset. One of the important biases is that the NA-ADNI participants are primarily enrolled in Caucasian populations[21]. Because differences in cortical structure have been reported between Caucasian and East Asian adults[22], the performance of AI prediction models trained using the NA-ADNI should be examined using other cohort datasets. To the best of our knowledge, no such evaluations using cross-cohort datasets have been conducted.

The present study focuses on solving the above two problems based on information that can easily be obtained with relatively low invasiveness and cost, such as magnetic resonance imaging (MRI)

[1]Imaging Technology Center, FUJIFILM Corporation, Kanagawa, Japan. [2]Department of Information Medicine, National Institute of Neuroscience, National Center of Neurology and Psychiatry, Tokyo, Japan. [3]Department of Psychiatry, Yokohama City University School of Medicine, Yokohama, Japan. [4]Department of Behavioral Medicine, National Institute of Mental Health, National Center of Neurology and Psychiatry, Tokyo, Japan. [5]Department of Clinical Epidemiology, Translational Medical Center, National Center of Neurology and Psychiatry, Tokyo, Japan. [6]Endowed Course for Health System Innovation, Keio University School of Medicine, Tokyo, Japan. *A list of authors and their affiliations appears at the end of the paper. ✉email: caihua.wang@fujifilm.com; yuanzhong.li@fujifilm.com

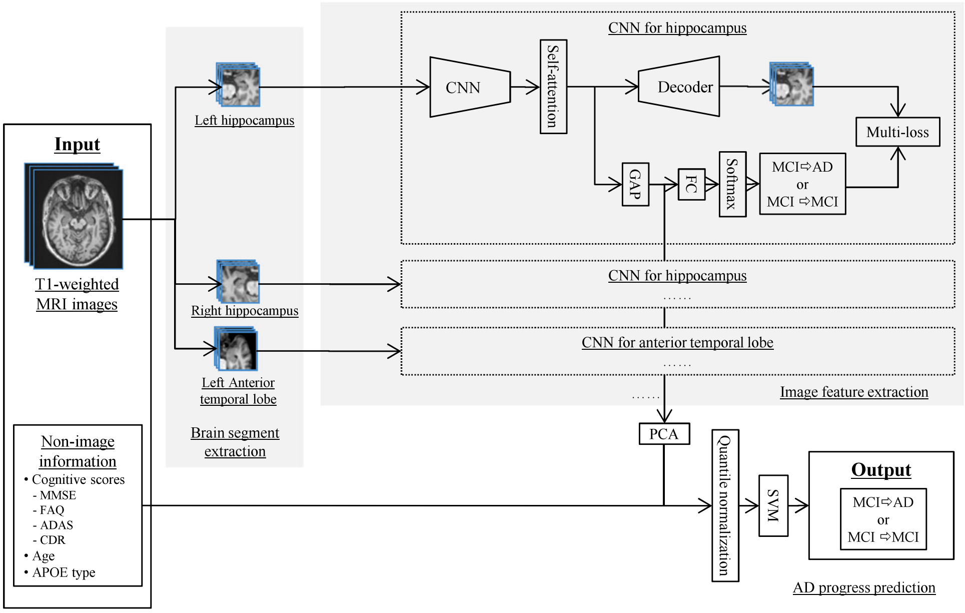

Fig. 1 A machine learning framework for predicting the progression of Alzheimer's disease. Multiple CNN-based image feature extractors with the same structure were trained on each brain segment, that is, the hippocampus and anterior temporal lobe in our study, to extract image features automatically. The CNN-based image feature extractor consists of a multi-layer convolutional neural network (CNN) block, a self-attention (SA) block, and an auto-encoder (AE) block. The feature vectors after the global average pooling (GAP) layer that encodes image features are passed to the PCA for dimensionality reduction. Finally, a linear SVM uses the low dimensional image features together with non-image information to make a final prediction.

images, which are used as brain imaging biomarkers; information such as positron emission tomography (PET) images is not adopted.

Here, we propose a hybrid machine learning framework (Fig. 1) that consists of an image feature extraction part (IFEP) that automatically extracts image features from predefined brain sub-regions related to cognitive decline, and an AD progression prediction part (APPP), which predicts whether a patient with MCI will convert to AD using extracted image features combined with clinical information. In the IFEP, a convolutional neural network (CNN)[23], which is a powerful tool used to extract features from images automatically[24], was applied. In the APPP, a linear support vector machine (SVM) classifier was used to make the final predictions. In the IFEP, to suppress the overfitting of deep learning, we introduced prior brain imaging studies[25–27] to use selected brain segments that have high correlations with cognitive decline to reduce the input image dimensions. To improve the performance and generalizability of our CNNs, mechanisms such as data augmentation[28], dropout[18], self-attention (SA)[29], and auto-encoding (AE)[30,31] were also used. A linear SVM tends to have higher generalizability compared with nonlinear classifiers[32], and using it as a classifier for the image features extracted by CNNs has been shown to be effective for classification[33]. Therefore, in the APPP, we selected it for the final predictions using both the image features extracted in the IFEP and the non-image information, including cognitive scores, age, and genetic apolipoprotein E (APOE) types.

The contribution of this study is that we investigated a hybrid approach consisting of a deep CNN for automatic image feature extraction from cognitive decline-related segments and a linear SVM for final prediction using both automatically extracted image features and non-image information to tackle the problem of predicting MCI progress, which has traditionally been difficult in machine learning because of the lack of sufficient training data. Our study also showed that such a hybrid approach could achieve prediction accuracies of 88% (area under the curve [AUC]: 95%) for the NA-ADNI and 84% (AUC: 91%) for the J-ADNI, thereby outperforming previous works. The most important point of our study is that such a hybrid approach can be more practically applicable for predicting the progress of MCI because of its high generalizability for different cohorts.

RESULTS

Training, validation, and test settings

The training, validation, and test datasets are shown in Fig. 2. Repeated 10-fold stratified cross-validation and test was adopted for evaluation of our model on the NA-ADNI dataset, and test was only carried out on the J-ADNI dataset that was completely unknown to the model. In the tenfold stratified cross-validation and test, the whole dataset was randomly split and stratified by class labels into 10 subsets. In each cross-validation and test, eight subsets were used for training; the remaining two were used for validation and test, respectively. The cross-validation and test were repeated 10 times by shifting the start subset of cross-validation and test setting to the next subset, so that each sample of all the datasets was used for test only once. The mean accuracy for all test subsets was calculated to evaluate our model in regard to the NA-ADNI dataset. For the evaluation of the J-ADNI dataset, the 10 models obtained in the above repeated 10-fold cross-validation and test were used, and the mean output probabilities of the models were used to classify each patient as sMCI or pMCI with a fixed threshold of 0.5.

Comparison with previous studies

A comparison of our results with those from previous studies is shown in Table 1. Because the same NA-ADNI dataset was used, the prediction performance was basically comparable, despite differences such as the follow-up periods used to define sMCI and pMCI, input information, and evaluation methods (validation, k-fold cross-validation, and k-fold cross-validation and test). For the case of using only brain images as input, our model (**M2** in Supplementary Table 2), based on only T1-weighted MRI images (accuracy: 78%, AUC: 85%, sensitivity: 78%, specificity: 78%), performed better than a previous deep neural network (DNN)[8] (accuracy: 75%, AUC: not available [NA], sensitivity: 75%, specificity: 75%), equivalent to a DNN[9] that used a much easier

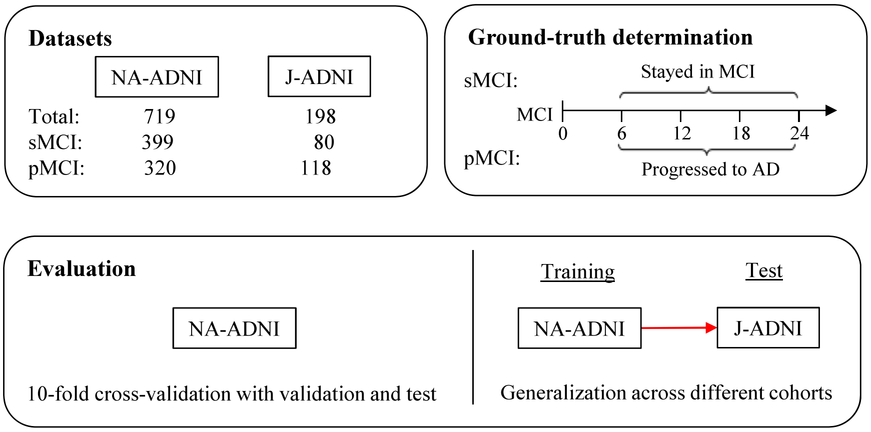

Fig. 2 Training, validation, and test datasets. We extracted data for training, validation, and test from the longitudinal datasets ADNI1, ADNI-GO, and ADNI2. If a participant was diagnosed as MCI and both the MRI T1-weighted images and other non-image information used in our model were available, we selected the related data as a candidate for this study. We evaluated our model in the following two ways: first, as in previous studies, we tested the model trained based on the NA-ADNI dataset using repeated 10-fold cross-validation and test; in each iteration, 80% of the samples were used for training, 10% for validation, and 10% for test; second, to evaluate generalization across cohorts, we tested the model, which was trained based on the ADNI dataset using a totally "unknown" J-ADNI dataset.

Table 1. Comparison with previous studies.

Method	Dataset	Input	Period	Accuracy	AUC	Sensitivity	Specificity
DNN [7]	NA-ADNI	MRI, FDG-PET	3 years	0.83	–	0.80	0.84
DNN [8]	NA-ADNI[a]	MRI	3 years	0.75	–	0.75	0.75
DNN [9]	NA-ADNI	MRI	1.5 years[b]	0.79	–	0.75	0.82
DNN [10]	NA-ADNI	MRI	3 years	0.83[c]	0.88[c]	0.76[c]	0.87[c]
SVM [11]	NA-ADNI	MRI, Cognitive scores	3 years	0.85	–	0.47	0.97
Random forest[12]	NA-ADNI	MRI, gender, age	1 year	0.79	–	0.82	0.74
SVM [13]	NA-ADNI	MRI, MMSE	3 years[b]	0.85[d]	0.90[d]	0.84[d]	0.88[d]
DNN [14]	NA-ADNI[a]	MRI, Cognitive scores, age	Series	–	0.76	–	–
DNN [15]	NA-ADNI	MRI, Cognitive scores, CSF, demographics	1 year	0.81	–	0.84	0.80
DNN [16]	NA-ADNI	MRI, Cognitive scores, APOE, gender, age	1 year	0.86[e]	–	0.82[e]	0.88[e]
Random forest[17]	NA-ADNI	MRI, PET, Cognitive scores, APOE	3 years	0.87[f]	0.87[f]	0.86[f]	–
Proposed (M2)	NA-ADNI	MRI	2 years	0.78	0.85	0.78	0.78
Proposed (M5)	NA-ADNI	MRI, Cognitive scores, APOE, age	2 years	0.88	0.95	0.88	0.88

[a] Relatively small datasets compared with the NA-ADNI were also used for both training and evaluation.
[b] Follow-up periods used for sMCI are longer than that of pMCI (5 years in DNN [9] and 4 years in SVM [13]).
[c] Trained with mixed groups of NC+sMCI and pSMI+AD.
[d] twofold validation accuracy.
[e] Tenfold cross-validation accuracy.
[f] Tenfold cross-validation accuracy, much higher accuracy of 91% on a small hold out test set.

task definition (accuracy: 79%, AUC: NA, sensitivity: 75%, specificity: 82%), but was inferior to a DNN[7] using MRI and PET images (accuracy: 83%, AUC: NA, sensitivity: 80%, specificity: 84%) and a DNN[10] using mixed groups of cognitively normal (CN) + sMCI and pSMI+AD groups for training (accuracy: 83%, AUC: 88%, sensitivity: 76%, specificity: 87%). For the case of using not only images, but also non-image information, the performance of our model (**M5** in Supplementary Table 2) (accuracy: 88%, AUC: 95%, sensitivity: 88%, specificity: 88%) was better than that of state-of-the-art models using SVM[13], DNN[16], and random forest[17] (accuracy: 85%–87%, AUC: 87%–90%). Considering that the state-of-the-art methods were evaluated only by validation datasets (not test datasets), the potential superiority of our model should be greater (see Supplementary Table 6). The experimental results indicated that our framework outperformed previous models and worked well on cross-cohort datasets, as shown in Table 1 and Fig. 3.

Cross-cohort evaluation results

Because only the NA-ADNI dataset obtained primarily using Caucasian populations was used for training our AI model, a non-Caucasian dataset was preferred for testing cross-cohort performance. The J-ADNI[34] is a multicenter, longitudinal observational study in Japan using an almost identical protocol to the NA-ADNI. As a result, J-ADNI participants are East Asian. By using J-ADNI as the test dataset, we evaluated the cross-cohort performance of our AI model. The receiver operating characteristic (ROC) curve evaluated using the J-ADNI is shown in Fig. 3, and the ROC curve of the NA-ADNI is also shown as a reference. The **M5** model (Supplementary Table 2) was used in both cases. Although the AUC of the J-ADNI was 0.91 and 4% lower than the NA-ADNI, the performance (Supplementary Table 3) was still comparable to those in the previous studies shown in Table 1, where the NA-ADNI was used as both a training and test dataset.

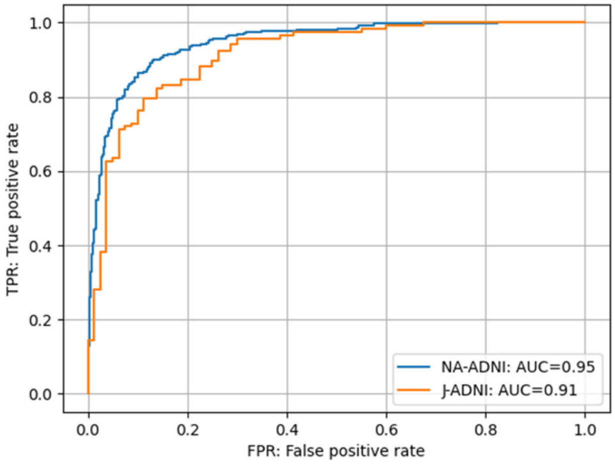

Fig. 3 Cross-cohort evaluation results. ROC curves for both the NA-ADNI and J-ADNI are shown. The AUCs of the NA-ADNI and J-ADNI are 0.95 and 0.91, respectively.

Visualization of our framework

First, to investigate which region the CNNs focused on in the IFEP, SmoothGrad[35] visualization was used. As shown in Supplementary Table 2, the performances using brain segments (**M2** and **M5**) were better than that using the whole brain (**M1 and M4**). The reason for this is shown in Fig. 4, where red represents a higher contribution ratio and blue represents a lower contribution ratio. The samples from S1 to S4 were selected from Fig. 5. Figure 4a shows that, even though **M1** and **M4** (Supplementary Table 2) focuses on the hippocampi and amygdalae, which play major roles in learning and memory, the variance of the "red" locations is quite large. On the other hand, as shown in Fig. 4b, **M2** and **M5** (Supplementary Table 1), which are based on brain segments, consistently focus on the hippocampi and amygdalae compared with the whole brain.

Second, to investigate the effectiveness of the features extracted by our model (**M5** in Supplementary Table 2), both the image and non-image features of the training samples were projected onto a two-dimensional (2D) space using UMAP[36] (Fig. 5). Because we wanted to see the global structure in the features, the value of the parameter "n_neighbors" was set to 50, which is bigger than the default value of 15. The input for UMAP involves 10-dimensional features obtained just after the "Quantile normalization" shown in Fig. 1. The 10 dimensions consist of four from the hippocampi and anterior temporal lobes (left and right), four from different cognitive scores, one from age, and one from APOE type. A blue "•" represents sMCI, which was correctly classified as sMCI by our model, and a blue "x" represents sMCI, which was wrongly classified as pMCI. A red "•" represents pMCI, which was correctly classified as pMCI, and a red "x" represents pMCI, which was wrongly classified as sMCI. Despite a small number of misclassification errors, most training samples were appropriately distributed, which suggests that effective features for classifying sMCI and pMCI were extracted successfully, and that a linear SVM is appropriate for classification tasks.

DISCUSSION

One main challenge of this study is how to extract meaningful features from images effectively by using a CNN with a limited number of training samples. CNNs are considered a powerful method that can automatically extract more effective features from images than manually designed. To realize this, CNNs require a huge dataset for training a large number of parameters. However, in the present study, while the number of available samples was relatively small (in the hundreds), the dimension of input images for each sample is huge (millions of voxels). In this case, a CNN can easily find a local optimal solution for its parameters. As a result, trained CNN models with local optimal solutions tend to lack the generalization of unknown datasets.

To extract robust image features from limited data, we introduced prior knowledge of brain imaging studies that investigated the correlation between brain segments and the progression of AD. In the IFEP, instead of one CNN for the whole brain, multiple CNNs were trained for different brain segments cropped from the whole brain images. Hippocampi and anterior temporal lobes (left and right) were finally selected experimentally. By adopting the brain segments, the dimension of input images for each sample was reduced to less than one-tenth of the whole brain. Moreover, each CNN was forced to learn image features from only the regions that were highly correlated with the progression of AD. The gradient maps shown in Fig. 4 demonstrate that the image features extracted by the CNNs of the cropped segments focus on more meaningful regions, such as hippocampi and amygdalae, compared with the whole brain. In our experiment, as shown in Supplementary Table 2, adding image features extracted from the whole brain (**M4**) improved the accuracy of only using non-image information (**M0**) by 2%, and using image segments (**M5**) instead of the whole brain improved the accuracy further by 1%.

The other challenge is what kind of classifier should be used to make a final prediction based on not only the features extracted in the IFEP but also the non-image information, including cognitive scores, age, and genetic APOE type. As described previously, the visualized feature distribution of sMCI and pMCI shown in Fig. 5 suggested that a linear classifier was appropriate. In the APPP, to restrain the overfitting induced by the limited data, instead of an E2E deep neural network, a linear SVM[32], which tends to have higher generalizability, was selected. The experimental results, in which the SVM model (**M5**, Supplementary Table 2) achieved 5% higher prediction accuracy than did the E2E model (**M3**, Supplementary Table 2), demonstrated the effectiveness of introducing the SVM.

In addition, our framework selected information that is relatively easy to be obtained as input. For example, even though PET imaging and cerebrospinal fluid (CSF) examinations are very useful for AD diagnosis, they were not selected because of their high cost and invasiveness. Furthermore, because the total processing time for predicting one case is about 30 seconds on a CPU-based PC, the hurdle to integrate our AI model into a practical clinical trial is low.

Successful and failure cases selected from the same UMAP visualization as that in Fig. 5 are shown in Supplementary Fig. 1. The "score" with a range from 0.0–1.0 is the output of the SVM; "0.0" and "1.0" represent the highest probability of sMCI and pMCI, respectively. The threshold for classification is 0.5. The samples S1–S6 shown in Supplementary Fig. 1a were selected from the cluster of sMCI samples, the boundary between the sMCI and pMCI clusters, and the cluster of pMCI samples.

S1 and S5 were selected from the cluster of sMCI samples. S1 is an sMCI case and was correctly classified as sMCI with a very low score of 0.15. S5 is a pMCI case and was classified wrongly as sMCI with a very low score of 0.09. Hippocampus atrophies of S5 are somewhat more progressive than S1, and cognitive scores are also slightly worse. However, because the score of S5 is lower than S1, our AI model shows that S5 has a lower probability to progress to AD. Note that the older population tends to have more progressive degradation of both hippocampus atrophies and cognitive scores caused by normal aging. The lower score of S5 might be affected by the fact that the age of S5 is 15 years older than that of S1.

S2 and S3 were selected from the boundary between the sMCI and pMCI clusters. S2 is an sMCI case and was classified correctly as sMCI with a low score of 0.30. S3 is a pMCI case and was classified correctly as pMCI with a score of 0.55, which is slightly higher than 0.5. Hippocampus atrophies of both S2 and S3 were unremarkable. On the other hand, the cognitive scores of S3 were slightly worse than those of S2. The difference in cognitive scores might have caused the different classification results.

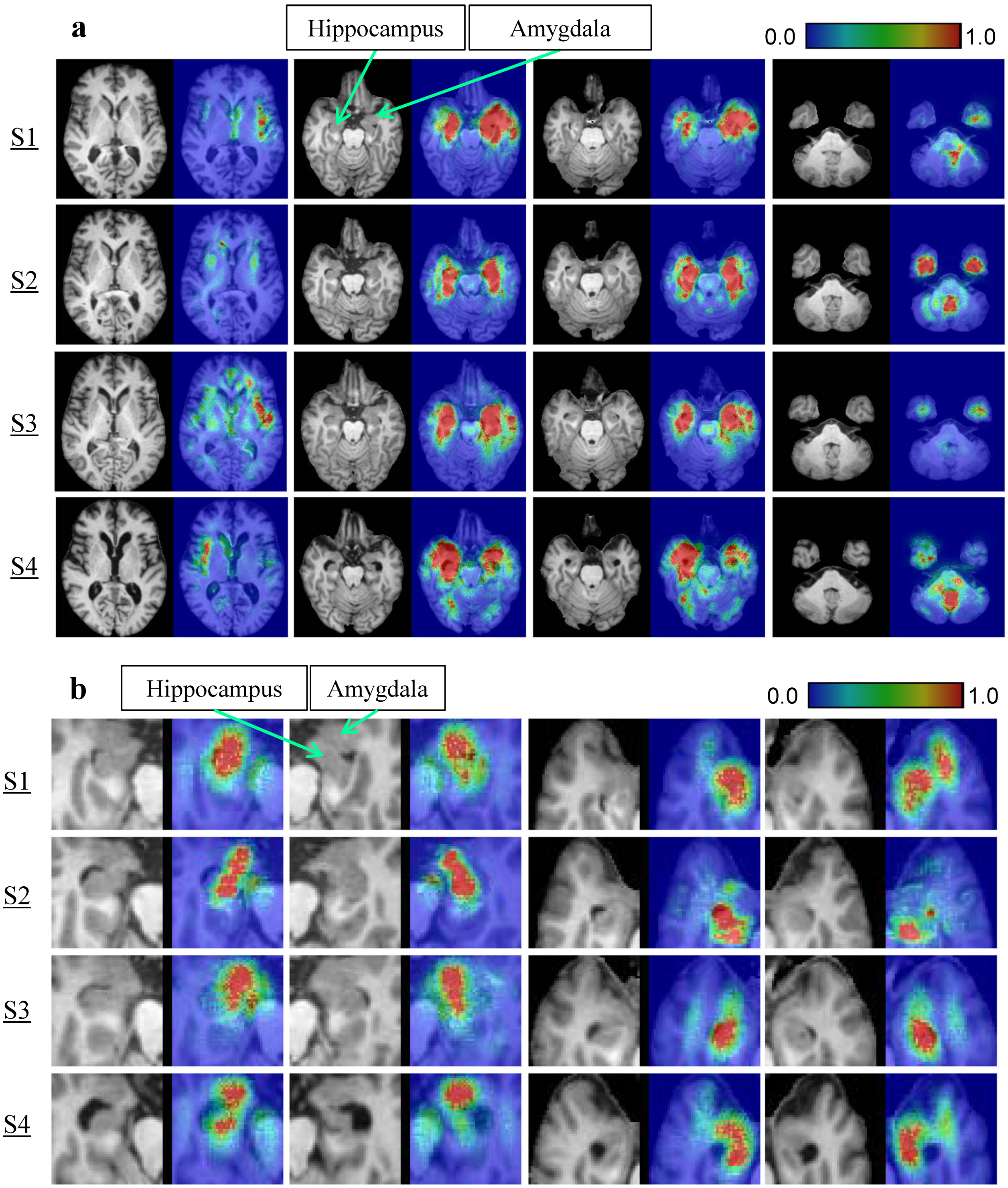

Fig. 4 Gradient maps of CNNs. a Gradient maps of CNNs of the whole brain. CNN visualization results of M1 (Supplementary Table 2) based on the whole brain. Red represents a high contribution ratio, and blue represents a low contribution ratio. **b** Gradient maps of CNNs of brain segments. CNN visualization results of M2 (Supplementary Table 2) based on brain segments. Red represents a high contribution ratio, and blue represents a low contribution ratio. Compared with the whole brain, the CNNs based on brain segments consistently focused on the hippocampi and amygdalae.

S4 and S6 were selected from the cluster of pMCI samples. S4 is a pMCI case and was classified correctly as pMCI with an extremely high score of 0.95. S6 is an sMCI case and was classified wrongly as pMCI, also with a high score of 0.97. Regarding hippocampus atrophy, S4 is a little progressive and S6 is very progressive. Regarding cognitive scores, both are poor. Both the obvious hippocampus atrophy and the poor cognitive scores indicate the reason S6 was misclassified as pMCI.

The strength of deep learning models, but also one of their vulnerabilities, is the ability to discern patterns in training data that humans cannot[37]. Sometimes, the vulnerability is striking, especially where there are biases in the training dataset. For

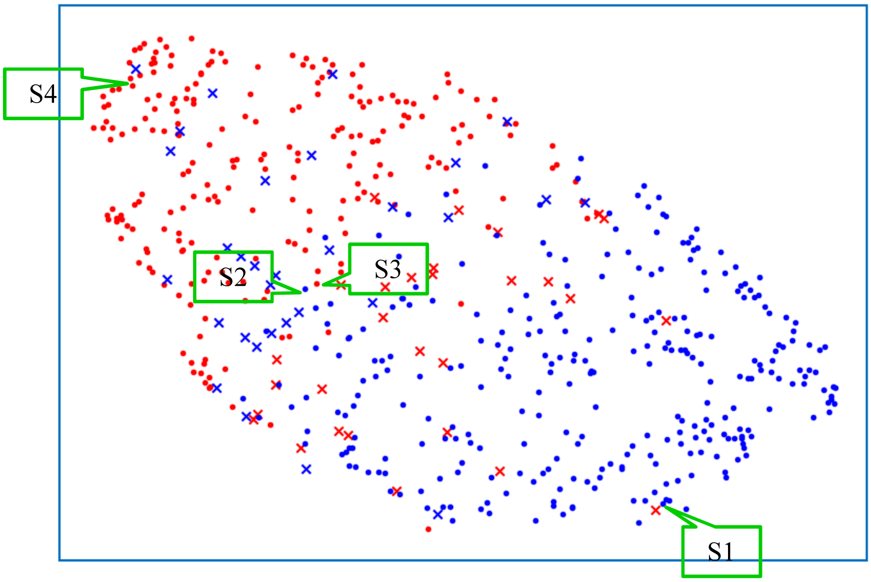

Fig. 5 Visualization by UMAP (10 ⇒ 2 dimensions). Training samples were plotted after the dimensions of their image and non-image features were reduced to 2 by using UMAP. Blue dots represent sMCIs that were correctly classified as sMCIs by our model, and blue "xs" represent sMCIs that were incorrectly classified as pMCIs. Red dots represent pMCIs that were correctly classified as pMCIs by our model, and red "xs" represent pMCIs that were classified incorrectly as sMCIs. Samples S1 to S4 were selected for CNN visualization in Fig. 4.

example, unintended effects caused by cohort (North America and East Asia) and gender biases have been reported in gender recognition systems[38]. As mentioned previously, because our AI model was trained based on the NA-ADNI dataset, which has a relatively small number of samples, mainly Caucasian, it is necessary to evaluate the trained model using a completely different dataset. In the present study, we evaluated its performance using the J-ADNI dataset, which consists of East Asian samples and was totally unknown to our AI model. Figure 3 shows that the performance of **M5** (Supplementary Table 2) using the J-ADNI. Furthermore, as shown in Table 1 and Supplementary Table 3, the prediction accuracy using the J-ADNI was 84% and comparable to previous studies that used the NA-ADNI as both the training and test datasets. The high performance shows the potential of our AI model to be incorporated in a practical clinical trial. A previous SVM study[13] (Table 1) also used the J-ADNI dataset for evaluation of their model trained on an NA-ADNI dataset, but the evaluation was only carried out for CN and pMCI subjects (accuracy: 91%, AUC: 97%), because their definition of sMCI required a follow-up period of over 4 years, but the longest follow-up period in the J-ADNI is 3 years.

Supplementary Table 4 shows the results of an ablation study of neural network architecture (that is, DenseNet with and without SA and AE). AE contributed the most in the NA-ADNI dataset, whereas SA contributed the most in the J-ADNI dataset. The final DenseNet model with both SA and AE showed stable performance for both the NA-ADNI and J-ADNI datasets.

Supplementary Table 5 shows the performance of an alternative model in which SVM linear kernel was replaced by radial basis function kernel. Compared with Supplementary Table 4, the two models were nearly equivalent, except that the AUCs of the linear model were slightly better, implying that the linear SVM is preferable.

Generally, an end-to-end deep learning classification mode often suffers from overfitting when the dataset is small, and replacing classification layers in the end-to-end model with a linear SVM can improve its robustness against overfitting. We investigated the difference in accuracies between the validation dataset, which was used to select the best parameters for a model, and the test dataset, to which the selected model was applied. The lower difference of the two accuracies suggests higher model generalizability. Supplementary Table 6 shows that the accuracy difference between the validation and test of hybrid model (**M5**) was much smaller than that of the end-to-end model (**M3**). A similar tendency was also seen in the models using images only (**M2** and **M6**). Supplementary Table 6 also shows that the accuracy varied widely among the randomly split validation and test subsets because the dataset was relatively small. This strongly implies the existence of bias when a small, randomly-selected evaluation subset is used, and repeated 10-fold stratified cross-validation and test is desirable to evaluate a developed model objectively.

The first limitation of this study is that, even though the ADNI dataset, which is the largest worldwide regarding AD, was used in this study, the size was still relatively small. Therefore, we proposed the hybrid model to suppress overfitting.

The second limitation is that, because we aimed at contributing to clinical trials, only the baseline data were used as input. To extend its application to common clinical diagnoses, where longitudinal data are usually available, there is room for improvement by utilizing not only the baseline data, but also the whole longitudinal data.

The third limitation is that our hybrid model showed superiority to end-to-end models using DenseNet as a backbone when the available dataset was limited. However, end-to-end models, including those using other state-of-the-art architectures such as vision transformer[39] as backbones, should be continuously investigated, because projects like the ADNI are ongoing and datasets will become larger.

In this paper, we proposed a hybrid machine learning framework to overcome the lack of training samples in the field of AD progression prediction. The experimental results showed that it worked well, even on different cohorts. Utilizing this framework to stratify patients with MCI who have higher risk of progression to AD might improve the success rate in future clinical trials of drugs for treating AD. To achieve this goal, the most important future tasks are to conduct retrospective evaluations based on previous trial data involving drug efficacy and to demonstrate the benefits of integrating AI quantitatively. In addition, as recent studies have shown that new biomarkers such as plasma phospho-tau are also effective for prediction[40], it is necessary to consider adding these biomarkers to input in future models.

METHODS

Datasets used in the study

The data used for training, cross-validation, and test in our study were obtained from the ADNI1, ADNI-GO, and ADNI2 datasets in the NA-ADNI database (http://adni.loni.usc.edu/ADNI). The NA-ADNI is a cohort study launched in 2003 as a public–private partnership, led by principal investigator Michael W. Weiner M.D., and carried out across 55 research centers in the USA and Canada. The clinical coordination center of NA-ADNI established a network of clinical sites and developed a plan for the recruitment and retention of subjects, and furthermore prepared the final clinical protocol and informed consent, which is distributed to the sites for local institutional review board (IRB) approval[6], Subjects were willing and able to undergo test procedures, including neuroimaging and follow-up, and written informed consent was obtained from participants[17]. Over 2000 participants with normal cognition and patients with MCI or AD were recruited for the present study. The first cohort, referred to as ADNI-1, consists of 800 individuals: 200 CN individuals, 400 with late MCI, and 200 with mild dementia. ADNI-GO, the second cohort, included about 200 additional individuals with early MCI. In ADNI-2, more participants at different stages of AD were recruited to monitor the progression of AD. ADNI-3, which is presently enrolling additional CN individuals and patients with MCI and dementia, was not included in our study, because no diagnosis information is currently available.

Longitudinal data for each participant were examined, and the samples used in our study were selected as follows. From the longitudinal data, we first identified candidate baseline points at which the participant was diagnosed as having MCI and both T1-weighted MRI images and other non-image information used in our model were available. The non-image information includes cognitive test scores on the Mini-Mental State Examination (MMSE), Functional Activities Questionnaire (FAQ), Clinical Dementia Rating (CDR) Sum of Boxes score, and Alzheimer's Disease Assessment Scale Cog-11 (ADAS), APOE type, and demographic information, such as gender and age. If the participant progressed to AD within 2 years from the baseline point, we labeled that participant, in association with the T1-weighted MRI images and non-image information at baseline, as a pMCI sample; otherwise, as an sMCI sample. There were 399 sMCI and 430 pMCI samples extracted for the training, validation, and test of our prediction model in this experiment.

The same procedure was applied to the J-ADNI dataset, which was collected at the 38 participating clinical sites using the same protocol as that for the ADNI, but from population of East Asian adults and in a shorter period (3 years). Approval for the J-ADNI study protocol was obtained from the local ethics committees or institutional review committees at the 38 participating clinical sites, including the principal investigator's site (The University of Tokyo). Informed written consent was obtained from all the participants at each clinical site[21]. In total, 80 stable and 118 progressive MCI samples were extracted from the J-ADNI dataset for the evaluation of our trained model.

Detailed information of the NA-ADNI and J-ADNI datasets is shown in Supplementary Table 7.

Image shape normalization

The 3D T1-weighted MRI images used in our study were first transformed into Montreal Neurological Institute (MNI)[41] space by aligning T1-weighted MRI images to atlas images (standard images) created in MNI space. The template images of the MNI152 NLIN 2009a atlas[42] were used. To align robustly and accurately T1-weighted MRI images, which were obtained under protocols used in NA-ADNI or J-ADNI studies, to the template images of the atlas, which were created in MNI space, a coarse-to-fine approach containing landmark-based and image registration-based alignment steps was developed.

In the landmark-based alignment step, six locations in the brain with distinct local anatomic structures were selected as landmarks to be used for robustly aligning T1-weighted MRI images to the atlas template images. Landmarks specified manually for the atlas template images are shown in Extended Fig. 2. These landmarks are the left and right eyes, body of fornix (front), midbrain (center), 4th ventricle (center), and corpus callosum (top). A region-based CNN (R-CNN)[43] was trained to detect the same landmarks from the T1-weighted MRI images. We trained the R-CNN with 290 cases collected from multiple datasets, including the NA-ADNI dataset, OASIS Brains dataset[44], and locally obtained data. Landmarks of training images were specified manually under the supervision of a radiological technologist.

When landmarks of T1-weighted MRI images were detected by the trained R-CNN, we set the center point of each landmark detected from each T1-weighted MRI image to correspond with that of the same landmark manually specified in the atlas template image. Using six pairs of corresponding central points of the landmarks in the T1-weighted MRI and atlas template images, we identified a linear transform, including 3D rotation, translation, and scaling, by minimizing the total error between the transformed central points of the landmarks detected from the T1-weighted MRI images and that of the atlas template images as follows:

$$\sum_{i=1}^{6} \| Q_i - (SRP_i + T) \| \overset{yields}{\rightarrow} min \tag{1}$$

where P_i and Q_i are the central points of the ith landmark in the atlas template and T1-weighted MRI images, respectively. S, R, and T denote 3D scaling, rotation, and translation, respectively. We used a linear transform with 9 degrees of freedom, called a similarity transform, instead of an affine transform because we only wanted to normalize the location, orientation, and size of the T1-weighted MRI images, not change the internal detail anatomic structures.

In the image registration-based alignment step, the initial similarity transform obtained from the landmark-based alignment was further refined by using an image registration technique that utilizes mutual information as an image similarity metric[45–47]. In this study, a brain mask of the atlas template images, which is available together with the atlas, was also used to ensure that the image registration was only performed on the brain region. The image registration technique used in this study was implemented based on the open-source Insight Tool Kit (ITK) image registration framework[48].

Image intensity normalization

As T1-weighted MRI images acquired in different sites with different equipment may have different biases in terms of intensity distribution, the normalization of image intensity is an important pre-procession for many image analysis tasks, and many methods have been proposed in the past[49–52]. In the present study, we adopted a similar example-based intensity normalization approach that uses patches in the T1-weighted MRI images and atlas template images, which contain patches of the acquired and tissue contrasts desired[50], but implemented it in a simpler way for robustness and computational simplicity.

Supplementary Fig. 3 shows an outline of our image intensity normalization algorithm. Using the similarity transform obtained above, the T1-weighted MRI images were transformed to shape-normalized images to align to the atlas template images. Then, an augmented example-based image intensity normalization was applied to the shape-normalized images, the details of which are described below.

To ensure that the corresponding patches in the shape-normalized and atlas template images used for intensity normalization contained the same tissue, we performed nonlinear image registration between the shape-normalized and atlas template images. Nonlinear image registration was also implemented based on ITK[48], and used mutual information as the image similarity metric and B-spline as the image transform[47]. Similar to what was done in the shape normalization, the brain mask of the atlas template image was also used to ensure that the nonlinear image registration was only performed on brain regions.

Using the B-spline transform obtained in the nonlinear registration, the shape-normalized images were transformed nonlinearly to the atlas template images, called the warped images, as follows:

$$I_w(X) = I_n(T_B(X; \varphi)) \tag{2}$$

where I_n and I_w denote the shape-normalized and warped images, respectively, and $T_B(X; \varphi)$ is the B-spline transform between the shape-normalized and atlas template images with parameters of φ.

In the warped images I_w, tissues in the normalized image I_n were nonlinearly deformed to align to those in the atlas template images. Using the brain mask of the atlas template images, we calculated the minimal box that contained the brain region, and divided the minimal boxes of both warped images I_w and atlas template images I_a into 8 × 8 × 8 blocks. For each pair of blocks in the warped and atlas template images at the same location, we calculated intensity histograms for both blocks of the pair and normalized the histogram range of the warped image block to that of the atlas template image block. Again, the brain mask of the atlas template image was used to exclude the voxels outside of the brain region based on the intensity histogram calculation. We normalized the histogram range instead of the histogram itself[52] because the portion of tissues of different types may change substantially among subjects, and normalizing the histogram itself could lead to over-normalization. On the other hand, the histogram range, the lower and upper bounds of which stand for the intensities of CSF and white matter, respectively, is more stable.

For each block B_k, let u_k^a and u_k^w be the upper bounds and l_k^a and l_k^w be the lower bounds of the intensity histogram of the block in the atlas template images I_a and the warped images I_w. We defined a linear intensity transform for block B_k as follows:

$$\acute{I}_k^w = \frac{u_k^a - l_k^a}{u_k^w - l_k^w} I_k^w - \frac{u_k^a - l_k^a}{u_k^w - l_k^w} l_k^w + l_k^a \equiv a_k I_k^w + b_k \quad (3)$$

where $I_k^W, k = 1, \cdots, N$ denotes the intensity of voxels in block k of the warped images I_w, and $\acute{I}_k^w$ is its normalized intensity.

To obtain a continuous intensity transform over the whole images, we regarded Eq. (3) as the intensity transform at the gravity center of each block B_k, that is, the gravity center of brain mask of block B_k, and then used a kernel function to expand a_k and b_k, two parameters that represent the intensity transform, to the whole image continuously. Using the gravity center $\overline{X}_k^w$ of block B_k and its intensity transform parameters a_k and b_k, we calculated two continuous parameter maps $\boldsymbol{a_w}(X^w)$ and $\boldsymbol{b_w}(X^w)$ of the intensity transform using a kernel function, defined as follows:

$$\boldsymbol{p_w}(X^w) = \sum_{k=1}^{N} p_k G\left(X, \overline{X}_k^w\right) / \sum_{k=1}^{n} G\left(X, \overline{X}_k^w\right) \quad (4)$$

where p stands for parameters a or b and X^w stands for the 3D coordinate of each voxel in the warped image I_w. $\mathrm{G}(\mathrm{X}, \overline{X})$ is defined as $\mathrm{G}(\mathrm{X}, \overline{X}) = \exp\left(-\frac{\|X - \overline{X}\|}{2\left(s_x^2 + s_y^2 + s_z^2\right)}\right)$, where s_x, s_y, and s_z are the block sizes. N is the number of blocks. Using the parameter map, we can normalize intensities in the warped image I_w voxel-wise.

As Eq. (4) was defined on the warped images I_w, to obtain the intensity transform on the shape-normalized images I_n, we back-projected the parameter map obtained on the warped images I_w to the shape-normalized images I_n using the inverse of the B-spline transform obtained in the nonlinear image registration.

$$\boldsymbol{p_n}(X^n) = \boldsymbol{p_w}\left(T_B^{-1}(X^n, \varphi)\right) \quad (5)$$

where X^n stands for the 3D coordinate of each voxel in the shape-normalized images I_n and p stands for parameters a or b. By applying the back-projected intensity transform parameters to the shape-normalized images I_n, we can obtain an intensity normalization.

Brain segment extraction

After the intensity normalization was carried out, a skull-stripping process was performed on the shape-normalized images to extract the brain region. To achieve this, we trained a V-net[53] with four layers. The V-net was trained using the same dataset as that used for the landmark detection. The ground truth of the brain regions of the training data was obtained by manual editing under the supervision of the radiological technologist. The segmentation results were generated automatically by an image registration method, which further included diffeomorphic image registration[54] implemented in ITK.

From the normalized (in shape and intensity) and skull-tripped image, brain segments of hippocampi (left and right) and anterior temporal lobes (left and right) were extracted. The locations of these segments were identified in the atlas template image manually and their sizes were fixed to 64 × 64 × 64 voxels, which was large enough to contain each segment of interest with a necessary margin in the normalized image. For each sample, four brain segments, which contained hippocampi and anterior temporal lobes, both left and right, at the same locations specified in the atlas template images, were extracted from its normalized image, respectively.

Image feature extraction

A densely connected convolutional network (DenseNet)[55], which has confirmed superiority over other network architectures for the same task on a different dataset[56], was adopted as the backbone of our architecture for image feature extraction. As shown in Supplementary Fig. 4, in our architecture, the DenseNet backbone was enhanced by adding an SA layer and an AE to improve its stability on limited training samples.

Supplementary Table 8 shows the detailed architecture of our DenseNet backbone, which is basically the same as DenseNet-121[55], where the 2D convolution and pooling layers were modified to 3D ones. The size of the first convolution layer was also changed to a smaller one because of the smaller size of the brain segment images.

Between the last dense block and the classification layer of the original DenseNet-121, an SA layer was inserted. While DenseNet is a method used to extract local relationships from images hierarchically, the SA highlights important positions in a feature map and helps to extract a global relationship. In our model, by adding an SA layer just after the DenseNet backbone, we could extract features representing both the local and global relationships.

Beside the classification task, in which an inputted brain segment was classified to sMCI or pMCI classes, another task of the AE, which recovers the imputed brain segment image from the image features, was added to the DenseNet backbone. By introducing the AE task, which was considered more robust than classification with a small number of samples, the robustness of the entire architecture was expected to be improved. For the decoder, we employed a simple five-layer network consisting of deconvolution and up-sampling layers, the details of which are shown in Supplementary Table 9.

To optimize the multi-task architecture, a mixed loss function of the two tasks was defined as follows:

$$\mathrm{Loss} = (1 - \alpha) L_{class} + \alpha L_{AE} \quad (6)$$

where L_{class} represents the classification error function for prediction defined by cross-entropy loss, L_{AE} denotes the AE error function defined by smooth L1 loss, and α is a hyper-parameter that controls the effect of the AE, which is set to 0.8 empirically.

The two networks with the same architecture shown in Supplementary Fig. 4 were trained for the brain segments of hippocampi and the anterior temporal lobe independently. Because the left and right segments of brain are basically symmetrical, images of the left hippocampus were flipped horizontally and then used with images of the right hippocampus to train the hippocampus network; this process was also applied to the anterior temporal lobe segment. The output of the global average polling of the trained network was extracted as image features and used with other non-image features for final classification, as described below.

The intensity of all images for training and testing were normalized to [0, 1] by a fixed maximum intensity of 400 in advance. Because the number of samples was small, we also augmented the training images before they were inputted to the network. The following augmentations were carried out for each training image:

1. A similarity transform with random rotations ranged from −2.0 to 2.0 degrees, scaling factors from 0.95 to 1.05, and translation from −4.0 to 4.0 voxels, around or along the x, y, and z axes, respectively.
2. A linear gray scale transform with a random slope from 0.95 to 1.05 and a random shift from −0.05 to 0.05, where the gray value of the images was normalized to [0, 1].

To avoid the influence of the outside border generated by the random similarity transform augmentation, the augmented images were cropped into a 48 × 48 × 48 voxel size, and the cropped images were used as the input for our network.

Linear SVM classification

Because the number of training samples is limited in the machine learning field of AD progression prediction, we adopted a linear SVM instead of E2E deep learning as the final predictor to improve generalizability. After a 128-dimensional image feature was extracted from each brain segment by CNNs, principal component analysis was used to reduce the dimension to 1 for each segment. We also attempted to reduce the dimension to 2 or 3 for each brain segment, but there were almost no changes in accuracy. Then, non-image information, including cognitive scores (MMSE, FAQ, CDR, and ADAS), age, and APOE type, was combined with the reduced image features of the four segments as the SVM input. Here, we transformed APOE type to a value representing AD risk based on the following rules:[57] for ε2/ε3, value = 0.6; for ε3/ε3, value = 1.0; for ε2/ε4 and ε4/ε2, value = 3.2; and for ε4/ε4, value = 11.6. Because the distributions of these multimodal inputs differ greatly, quartile normalization[58] is used to normalize the combined features before the SVM is applied. For validation and test, the same quartile normalization used for training was applied to the combined features.

End-to-end (E2E) classification

We also implemented two E2E models using the same backbone architecture shown in Supplementary Fig. 4; one uses only images for prediction and the other uses both images and non-image information, as shown in Supplementary Fig. 5. For the model that uses both image and non-image information, bilinear fusion[16] was used to combine the image features extracted from images and non-image features extracted from non-image information. The same AE as shown in Supplementary Fig. 4 was also adopted, but this was omitted because of space limitations. In the architecture, two image feature extraction modules were configured in parallel to extract image features from the images of the hippocampi and anterior temporal segments, and the extracted image features were concatenated together. For the model using both image and non-image information, the non-image features were first embedded and then extended to the same dimension of the image features. Bilinear fusion was then applied to the image features and extended non-image features to generate the multimodal features for classification.

In our experiments, although the E2E model achieved high validation accuracy, a relatively large decline in test accuracy (from 87% to 83%) was observed. On the other hand, only a slight decline in test accuracy was seen (from 89% to 88%) for our proposed architecture, which combined image and non-image features with a linear SVM, while it achieved the highest validation accuracy. This observation may imply that a simpler combination of multimodal features may be more stable than an E2E deep learning approach when the number of samples available is relatively small.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used for model training, validation, and test are publicly available at the following URLs:
1 NA-ADNI dataset: http://adni.loni.usc.edu/.
2 J-ADNI dataset: https://humandbs.biosciencedbc.jp/en/hum0043-v1.

CODE AVAILABILITY

The DenseNet code, which was used as the backbone of our architecture, is available at https://github.com/liuzhuang13/DenseNet. The Faster R-CNN code used for brain landmark detection is available at https://github.com/rbgirshick/py-faster-rcnn. The Insight Toolkit: ITK, on which our image registration algorithm was implemented, is available at https://itk.org.

Received: 17 August 2021; Accepted: 11 February 2022;
Published online: 12 April 2022

REFERENCES

1. Alzheimer's Association. Alzheimer's disease facts and figure. *Alzheimer's Dement.* **15**, 321–387 (2019). 2019.
2. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimer's Dement.* **5**, 272–293 (2019).
3. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2020. *Alzheimer's Dement.* **23**, 1–24 (2021).
4. Manly, J. J. et al. Frequency and course of mild cognitive impairment in a multiethnic community. *Ann. Neurol.* **63**, 494–506 (2008).
5. Petersen, R. C. et al. Practice guideline update summary: mild cognitive impairment. *Neurology* **16**, 126–135 (2018).
6. Mueller, S. G. et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. North Am.* **15**, 869–877 (2005).
7. Lu, D., Popuri, K., Ding, G. W., Balachandar, R. & Beg, M. F. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci. Rep.* **8**, https://doi.org/10.1038/s41598-018-22871-z (2018).
8. Basaia, S. et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clin.* **21**, https://doi.org/10.1016/j.nicl.2018.101645 (2019).
9. Zhang, J. et al. A 3D densely connected convolution neural network with connection-wise attention mechanism for Alzheimer's disease classification. *Magn. Reson. Imaging* **78**, 119–126 (2021).
10. Abrol, A. et al. Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease. *J. Neurosci. Methods* **339**, https://doi.org/10.1016/j.jneumeth.2020.108701 (2020).
11. Tam, A. et al. A highly predictive signature of cognition and brain atrophy for progression to Alzheimer's dementia. *GigaScience* **8**, https://doi.org/10.1093/gigascience/giz055 (2019).
12. Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A. & Rueckert, D. Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. *Sci. Rep.* **8**, https://doi.org/10.1038/s41598-018-29295-9 (2018).
13. Syaifullah A. et al. Machine learning for diagnosis of AD and prediction of MCI progression from brain MRI using brain anatomical analysis using diffeomorphic deformation. *Front. Neurosci.* **11**, https://doi.org/10.3389/fneur.2020.576029 (2021).
14. Nakagawa, T. et al. Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI images. *Brain Commun.* **2**, https://doi.org/10.1093/braincomms/fcaa057 (2020).
15. Lee, G., Nho, K., Kang, B., Sohn, K. A. & Kim, D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci. Rep.* **9**, https://doi.org/10.1038/s41598-018-37769-z (2019).
16. Goto, T., Wang, C., Li, Y. & Tsuboshita, Y. Multi-modal deep learning for predicting progression of Alzheimer's disease using bi-linear shake fusion. *Proc. SPIE (Med. Imaging)* **11314**, 452–457 (2020).
17. El-Sappagh, S. et al. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci. Rep.* **11**, https://doi.org/10.1038/s41598-021-82098-3 (2021).
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
19. Wang, X. et al. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 3462-3471 (2017).
20. Deng, J. et al. ImageNet: A large-scale hierarchical image database. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 248-255 (2009).
21. Iwatsubo, T. et al. Japanese and North American Alzheimer's disease neuroimaging initiative studies: harmonization for international trials. *Alzheimer's Dement.* **14**, 1077–1086 (2018).
22. Kang, D. W. et al. Differences in cortical structure between cognitively normal East Asian and Caucasian older adults: a surface-based morphometry study. *Sci. Rep.* **10**, https://doi.org/10.1038/s41598-020-77848-8 (2020).
23. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
24. Yamaguchi, H. et al. Three-dimensional convolutional autoencoder extracts features of structural brain images with a "diagnostic label-free" approach: application to Schizophrenia Datasets. *Front. Neurosci.* **07**, https://doi.org/10.3389/fnins.2021.652987 (2021).
25. Wang, L. et al. Alterations in cortical thickness and white matter integrity in mild cognitive impairment measured by whole-brain cortical thickness mapping and diffusion tensor imaging. *Am. J. Neuroradiol.* **30**, 893–899 (2009).
26. Misra, C., Fan, Y. & Davatzikos, C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *Neuroimage* **44**, 1415–1422 (2009).
27. Risacher, S. L. et al. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* **6**, 347–361 (2009).
28. Shorten, C. & Khoshgoftaar, T. A survey on image data augmentation for deep learning, *J. Big Data*, **6**, 1–48, (2019).
29. Vaswani, A. et al. Attention is all you need. *Proc. Int'l Conf. on Neural Information Processing Systems*, 6000–6010 (2017).
30. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P. A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
31. Hinton, G. E., Krizhevsky, A. & Wang, S. D. Transforming auto-encoders. *Proc. Int'l Conf. on Artificial Neural Netw.* 44-51 (2011).
32. Brereton, R. G. & Lloyd, G. R. Support Vector Machines for classification and regression. *The Analyst Online* **135**, https://doi.org/10.1039/B918972F (2010).
33. Girshick, R., Donahue J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 580-587 (2014).
34. Iwatsubo, T. Japanese Alzheimer's disease neuroimaging initiative: present status and future. *Alzheimer's Dement.* **6**, 297–299 (2010).
35. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. SmoothGrad: removing noise by adding noise, *Proc. Int'l Conf. on Machine Learning (ICML)*, https://doi.org/10.48550/arXiv.1706.03825 (2017).
36. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861. https://doi.org/10.21105/joss.00861 (2018).
37. Rajkomar, A., Dean, J. & Kohane, I. Machine learning in medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
38. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* **81**, 1–15 (2018).
39. Dosovitskiy, A. An image is worth 16x16 words: transformers for image recognition at scale. *Proc. Int'l Conf. on Learning Representations (ICLR)*, https://arxiv.org/abs/2010.11929 (2021).
40. Palmqvist, S. et al. Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. *Nat. Med.* **27**, 1034–1042 (2021).

41. Brett, M., Johnsrude, I. S. & Owen, A. M. The problem of functional localization in the human brain. *Nat. Rev. Neurosci.* **3**, 243–249 (2002).
42. Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R. & Collins, D. L. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, 539–541 (2009).
43. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1137-1149 (2016).
44. Marcus, D. S. et al. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* **19**, 1498–1507 (2007).
45. Maes, F., Vandermeulen, D. & Suetens, P. Medical image registration using mutual information. *Proc. IEEE* **91**, 1699–1722 (2003).
46. Oliveira, P. P. M. & Tavares, M. R. R. Medical image registration: a review. *Computer Methods Biomech. Biomed. Eng.* **17**, 73–93 (2014).
47. Mattes, D., Haynor, D. R., Vesselle, H., Lewellen, T. K. & Eubank, W. Nonrigid multimodality image registration. *Proc. SPIE 4322, Medical Imaging: Image Processing*, 1609-1620(2001).
48. Avants, B. B. et al. The Insight ToolKit image registration framework. *Front. Neuroinformatics* **8**, 1–13 (2014).
49. Reinhold, J.C., Dewey, B.E., Carass, A. & Prince, J.L. Evaluating the impact of intensity normalization on MR image synthesis. *Proc. SPIE 10949, Medical Imaging: Image Processing*, https://doi.org/10.1117/12.2513089 (2019).
50. Roy, S. & Carass, A. Magnetic resonance image example-based contrast synthesis. *IEEE Trans. Med. Imaging* **32**, 2348–2363 (2013).
51. Shinohara, R. T. et al. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage: Clin.* **6**, 9–19 (2014).
52. Sun, X. et al. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *BioMedical Eng. OnLine* 14, https://doi.org/10.1186/s12938-015-0064-y (2015).
53. Milletari, F., Navab, N. & Ahmadi, S. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Fourth Int'l Conf. on 3D Vision (3DV)*, 565-571 (2016).
54. Ashburner, J. A fast diffeomorphic image registration algorithm. *NeuroImage* **38**, 95–113 (2007).
55. Huang, G., Liu, Z. & Maaten, L. Densely Connected Convolutional Networks. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 4700-4708 (2017).
56. Islam, J. & Zhang, Y. Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks. *Brain Informatics* **5**, https://doi.org/10.1186/s40708-018-0080-3 (2018).
57. Hsiunga, G-Y.R & Sadovnick, A.D. Genetics and dementia: risk factors, diagnosis, and management. *Alzheimer's Dement.* **3**, 418–427 (2007).
58. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

ACKNOWLEDGEMENTS

This research was supported by Program on Open Innovation Platform with Enterprises, Research Institute and Academia, Japan Science and Technology Agency (JST, OPERA, JPMJOP1842). We acknowledge all members of the imaging technology center of Fujifilm Corporation for their helpful discussions. ADNI data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceuticl Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. ADNI investigators contributed to the design and implementation of the ADNI database and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

AUTHOR CONTRIBUTIONS

C.W. and Y.L. conceived the study. C.W., Y.T., T.G., and Y.L. implemented the experiments. T.S. contributed research materials and analytic tools. H.Y., Y.Y., A.S., and H.T. contributed to the definition of the progression of Alzheimer's disease and the analysis of the results. C.W. and Y.L. wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00577-x.

Correspondence and requests for materials should be addressed to Caihua Wang or Yuanzhong Li.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

Caihua Wang[7], Yuanzhong Li[7] and Tsubasa Goto[7]

[7]Imaging Technology Center, FUJIFILM Corporation, Kanagawa, Japan.