European Journal of Nuclear Medicine and Molecular Imaging
https://doi.org/10.1007/s00259-019-04676-y

ORIGINAL ARTICLE

Slice-selective learning for Alzheimer's disease classification using a generative adversarial network: a feasibility study of external validation

Han Woong Kim[1] · Ha Eun Lee[1] · Sangwon Lee[2] · Kyeong Taek Oh[1] · Mijin Yun[2] · Sun Kook Yoo[1]

Received: 18 August 2019 / Accepted: 27 December 2019

Abstract

Purpose The aim of this feasibility study was to use slice selective learning using a Generative Adversarial Network for external validation. We aimed to build a model less sensitive to PET imaging acquisition environment, since differences in environments negatively influence network performance. To investigate the slice performance, each slice evaluation was performed.

Methods We trained our model using a 18F-fluorodeoxyglucose ([^{18}F]FDG) PET/CT dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and tested the model with a Severance Hospital dataset. We applied slice selective learning to reduce computational cost and to extract unbiased features. We extracted features of Alzheimer's disease (AD) and normal cognitive (NC) condition using a Boundary Equilibrium Generative Adversarial Network (BEGAN) for stable convergence. Then, we utilized these features to train a support vector machine (SVM) classifier to distinguish AD from NC.

Results The slice range that covered the posterior cingulate cortex (PCC) using double slices showed the best performance. The accuracy, sensitivity, and specificity of our proposed network was 94.33%, 91.78%, and 97.06% using the Severance dataset and 94.82%, 92.11%, and 97.45% using the ADNI dataset. The performance on the two independent datasets showed no statistical difference ($p > 0.05$). Moreover, there was a statistical difference in the performance between using two slices and one slice as input ($p < 0.05$).

Conclusions Our model learned the generalized features of AD and NC for external validation when appropriate slices were selected. This study showed the feasibility of this model with consistent performance when tested using datasets acquired from a variety of image-acquisition environments.

Keywords Alzheimer's disease · [^{18}F] FDG PET/CT · Generative Adversarial Network · External validation · Feasibility study

Introduction

Alzheimer's disease (AD) is a degenerative brain disorder characterized by a decline in cognitive function. It mostly affects older people (> 65 years), so that the prevalence of AD has been sharply increasing with the rapid growth of the elderly proportion. Although it is currently difficult to cure AD, especially in advanced cases, early diagnosis of AD before the symptoms become severe would provide windows for effective treatment [1]. For the clinical diagnosis of AD, 18F-fluorodeoxyglucose ([^{18}F]FDG) positron emission tomography/computed tomography (PET/CT) is one of the most useful modalities. The imaging method called FDG PET/CT allows visualization of the glucose metabolism in the brain with high sensitivity and specificity over the course of diseases. It has been reported that there is reduction of glucose metabolism in the temporal, parietal, and posterior

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence)

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-019-04676-y) contains supplementary material, which is available to authorized users.

✉ Sun Kook Yoo
SUNKYOO@yuhs.ac

✉ Mijin Yun
YUNMIJIN@yuhs.ac

[1] Department of Medical Engineering, Yonsei University College of Medicine, Seoul, Republic of Korea

[2] Department of Nuclear Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

Published online: 24 January 2020

cingulate cortex (PCC) regions in patients with AD. The measurement of changes in the glucose metabolism help diagnose AD before the appearance of symptoms [2].

Recently, there have been a number of studies in which deep learning was applied in imaging analysis [3]. Deep neural networks, one of the machine learning methods, are used to solve problems in the field of recent image recognition that have not been solvable using conventional machine learning algorithms. Krizhevsky et al. showed an error rate of 15.3% in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012, which is lower than the 26.2% rate of a conventional machine learning method [4].

Deep learning methods have been studied in the medical imaging field for the classification of AD. Studies of AD classification using deep learning models include those using autoencoders [5, 6]. Liu et al. [5] applied several autoencoders with multi-layered neural network to combine multimodal features for AD classification. Suk et al. [6] designed stacked autoencoders to extract high-level features of multimodal ROI, and an SVM classifier was used to combine and classify AD. Moreover, CNN methods were applied to capture spatial features from PET or MRI scans [7, 8]. Glozman et al. [7] chose center slices of axial, coronal, and sagittal scans and applied 2D CNN to these. Then, the features from different views were concatenated and used for AD classification. Farooq et al. [8] used skull stripping and gray matter segmentation using MRI volume data. The slices were used as CNN model input where gray matter information existed. Moreover, 3D CNN was applied to take full advantage of 3D volume data [9, 10]. Hosseini et al. [9] applied 3D CNN to capture anatomical variations related to AD, such as hippocampus volume and brain volume. Liu et al. [10] used the entire PET volume to encode spatial volume features using 3D CNN.

Although previous studies have shown effective AD classification methods, limitations remained. Medical images are affected by the image acquisition environment. For example, the partial volume effect (PVE) depends on the imaging acquisition system [11]. Specific areas are blurred in different resolution for each dataset. Thus, performance may be reduced, which leads to poor clinical utility. In addition, input slices from volume data need to be carefully chosen, since the performance differs on each slice. However, Glozman et al. [7] extracted center slices of the volume data. Farooq et al. [8] excluded slices that did not have gray matter information. Hence, the specific standards for selecting slices need to be set.

Therefore, this paper is to propose a model that applied slice selective learning using a BEGAN-based model to solve the above limitations. We trained our model with an ADNI dataset, then performed external validation with our Severance hospital dataset.

Methods

Dataset

For training and validation, we used [^{18}F] FDG PET/CT data selected from the AD Neuroimaging Initiative (ADNI) database. The ADNI open database for the diagnosis of AD contains medical imaging data such as MRI and PET, as well as biochemical biomarkers and other data. The ADNI dataset included data from 139 patients with AD and 347 NC participants. A dose of [^{18}F] FDG (185 MBq; 5 mCi) was injected into subjects. ADNI [^{18}F] FDG PET/CT images of six 5-min frames were obtained 30 to 60 min after the injection. The [^{18}F] FDG PET/CT images were reoriented into standardized voxels (1.5 × 1.5 × 1.5 mm).

For testing, we obtained an institutional dataset from the Nuclear Medicine Department of Severance Hospital (Seoul, South Korea). The Severance data were from 73 patients with AD and 68 NC participants. This [^{18}F] FDG PET/CT data collection was approved by the Institutional Review Board (4–2018-1010). All [^{18}F] FDG PET/CT images were acquired using a Discovery 600 (GE Medical Systems, Milwaukee, WI) PET/CT. Approximately 4.1 MBq of [^{18}F] FDG per kilogram of body weight was administered intravenously to the patients. Forty minutes after [^{18}F] FDG injection, PET images were acquired for 15 min. Spiral CT scans were performed for attenuation correction with 0.8 s rotation time, 60 mA, 120 kVp, 3.75 mm section thickness, 0.625 mm collimation, and 9.375 mm table feed per rotation. We reconstructed [^{18}F] FDG PET/CT images using the ordered subset expectation maximization algorithm (4 iterations and 32 subsets). The demographics of the ADNI dataset and Severance dataset are shown in Table 1.

Preprocessing

We processed the raw [^{18}F] FDG PET/CT scans using the method described by Jagust et al. [12]. Each [^{18}F] FDG PET scan was co-registered to the first frame of the raw [^{18}F] FDG PET/CT scan to reduce the effects of subject motion. We generated a single PET scan by averaging six dynamic frames and then reoriented the co-registered and averaged scan into a standard voxel image grid with 1.5 mm^3 voxels.

Next, we normalized the voxel intensity of the processed [^{18}F] FDG PET scans using an iterative method previously described. For the first iteration, the entire image was scaled to a mean intensity value of 1.0. Successive iterations masked voxels with intensity values lower than 0.5. We rescaled the remaining voxels to a mean of 1.0. The voxel intensity of the [^{18}F] FDG PET was normalized by repeating this process until the number of remaining voxels became constant.

After intensity normalization, each [^{18}F] FDG PET scan displayed the difference between a subject's brain size and

Table 1 Demographic information of ADNI dataset and Severance dataset

	ADNI		Severance	
	AD	NC	AD	NC
Number of patients	139	347	73	68
Gender(F/M)	91 F/48 M	175 F/172 M	44 F/29 M	38 F/30 M
Age	76.04 ± 7.86	76.31 ± 6.43	70.75 ± 9.51	63.54 ± 9.14
MMSE	27.06 ± 1.79	28.95 ± 1.43	21.67 ± 4.48	29.26 ± 0.99
CDR	0.47 ± 0.12	0.01 ± 0.06	0.72 ± 0.38	0.06 ± 0.16

Data are mean ± SD, *AD* Alzheimer's disease, *NC* normal control, *SD* standard deviation, *MMSE* mini-mental state examination, *CDR* clinical dementia rating

shape. We used a spatial normalization method based on the MNI-152 template [13]. Thus, the same brain regions appeared in the same position in all the patients' brain scans.

Slice selection

Due to insufficient data to train the model using 3D [^{18}F] FDG PET volume data, we reduced computational cost by using slice-selective learning. To extract regularized features for AD classification, we trained our model within the range from the amygdala to the end of the PCC. We chose this slice range because it covers places where the neuropathological or anatomical changes are known to take place in AD (e.g., the posterior cingulate cortex, hippocampus, and entorhinal cortex) [14–16]. Each coronal 2D slice was 1.5 mm thick, and we extracted at 3-slice intervals (4.5 mm) from the [^{18}F] FDG PET/CT datasets. Each slice was numbered from the back of the head. Generally, coronal slices were cut at 5–7.5 mm for metastases, infarcts, etc. in brain autopsy [17]. Our network was trained with two coronal slices for additional volume information.

Boundary equilibrium adversarial network

Goodfellow et al. [18] proposed Generative Adversarial Networks (GAN) containing generative and discriminative networks in which the adversarial training process of two networks could improve the performance of both networks. The generator synthesizes a fake image G(z), which is difficult for the discriminator to handle. At the same time, the discriminator is trained to distinguish whether the provided image is real or fake. However, in the original GAN, the training process is unstable because of the imbalance between the discriminator and the generator. Therefore, we applied the Boundary Equilibrium Generative Adversarial Network (BEGAN) structure to our network as proposed by Berthelot et al. [19]. The BEGAN architecture maintains equilibrium between the discriminator and generator.

Our model is a hybrid model of BEGAN with an added condition term proposed by Mirza et al. [20]. For the generator, we added a condition term to generate a corresponding image. We used one-hot vectors as labels of AD and NC. The discriminator was an autoencoder with a loss function derived from the Wasserstein distance. Unlike the discriminator of the original GAN, the discriminator was trained to maximize the Wasserstein distance between the real and fake image reconstruction loss, not the data distribution, whereas the generator was trained to minimize the distance. The reconstruction loss is defined as:

$$L(x) = |x - D(x)| \tag{1}$$

In the training process, the discriminator extracts features from real AD and NC. We utilized this latent feature vector in the bottleneck layer for AD classification. The objective function is defined as:

$$\begin{cases} L_D = |x - D(x)| - k_t|x - G(z|y)| \\ L_G = L(G(z|y)) \\ k_{t+1} = k_t + \lambda_k(\gamma L(x) - L(G(z|y))) \end{cases} \tag{2}$$

where x is a real input image, z is the noise input vector, and y is a label of AD or NC. Here, $k_t \in [0, 1]$ is a control parameter that maintains the equilibrium. The initial value is set to zero and updated as below. The term λ_k is the learning rate and is updated at each training step. The diversity ratio parameter $\gamma \in [0, 1]$ controls the tradeoff between image diversity and image quality. If γ gets low, the discriminator focuses more on reducing the reconstruction loss of the real image, which leads to higher quality of generated image with less diversity.

For convergence measurement, we derived a formula that determines whether the model has converged or diverged without visual inspection.

$$M_{global} = L(x) + \left|\gamma\, L(x) - L\Big(G(z|y)\right| \tag{3}$$

Network architecture

Our goal was to design a generalized enhanced model that could classify AD and NC with a variety of [^{18}F] FDG PET/CT institutional datasets. We propose a two-stage deep learning process, trained it using GAN, and then used the extracted features for AD classification using a fully connected layer and a support vector machine (SVM). The overview of our model is illustrated in Fig. 1.

The proposed network architecture for the generator and discriminator is shown in Fig. 2. The first stage was to train the GAN model. The generator $G : \mathbb{R}^{N_z} \rightarrow \mathbb{R}^{N_x}$ was designed to generate a fake image of 128 × 128. In the generator, noise vector z was randomly sampled from 64 dimensions of random uniform distribution $z \in [-1, 1]$. A condition term enabled our model to produce AD and NC $[^{18}F]$ FDG PET images. We used a 3 × 3 convolution process using exponential linear units (ELU). The generator maintains the number of filters as 128. The discriminator is the structure of the autoencoder. The encoder linearly increases the number of channels from 128 to 640, and the decoder uses the same architecture of the generator. We used the nearest neighbor interpolation method in 2×2 upsampling and applied 2×2 downsampling.

The second step was to train a classifier that consisted of one fully connected layer and SVM. Because the features from the discriminator were suitable for distinguishing AD and NC, the same features were used to train the classifier to classify AD and NC. We used the fully connected layer to reduce the feature dimensions from 81,920 to 1024. The encoded features were classified with the SVM using a linear kernel.

Statistical test

We compared our results using Pearson's chi-square test to show similarity in performance between the two different dataset results. *P* values were calculated for the slice range examined. All *p* values < 0.05 were considered statistically significant. Moreover, we compared the performance of our model one-slice input and two-slice input. We further validated our model with single slice, double slice, quadruple slice, and 3D volume data. We also compared the performance of our method with other AD classification methods. We used a receiver operating characteristic (ROC) curve to calculate the area under the curve (AUC) for performance comparison,

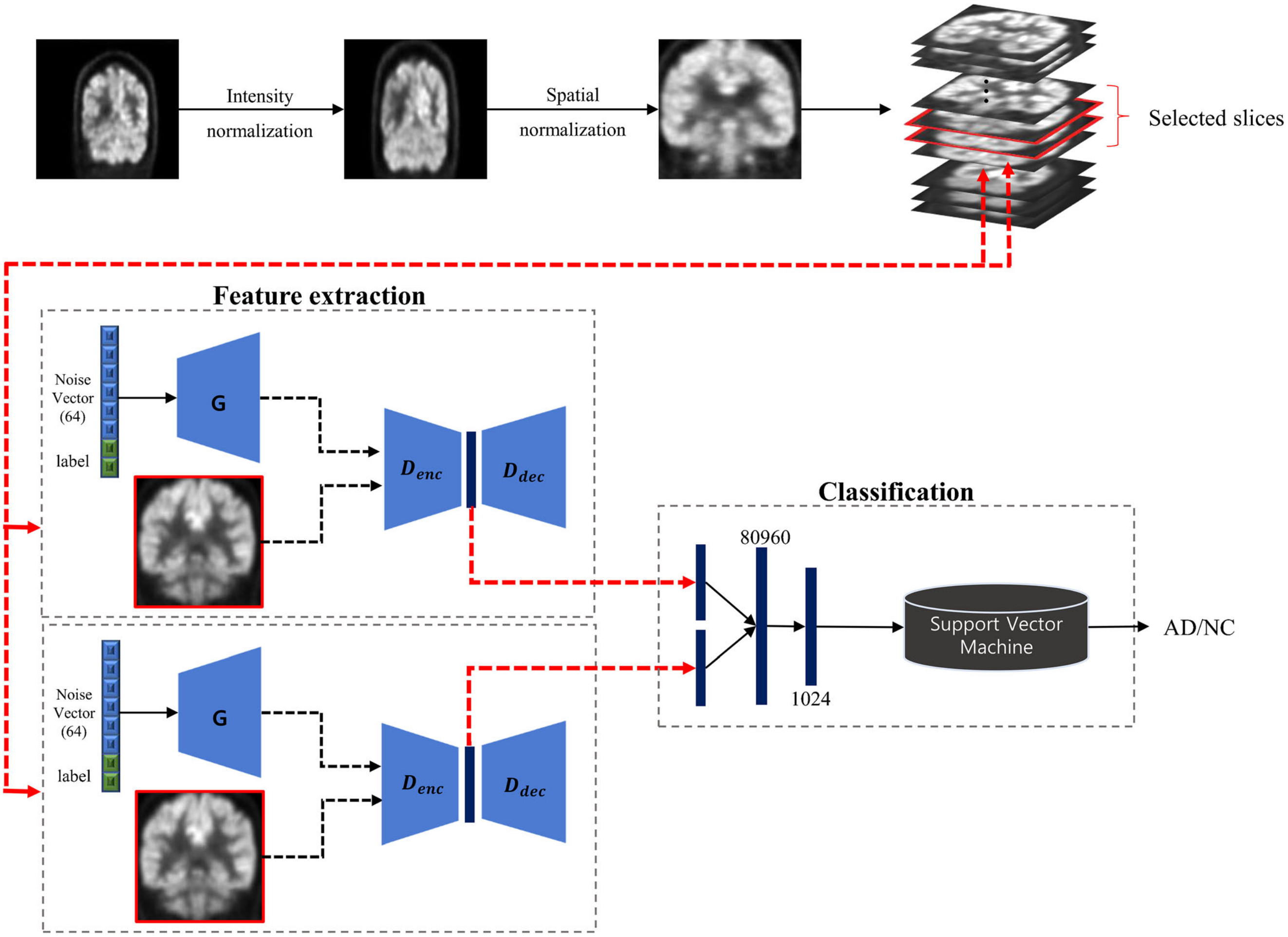

Fig. 1 Overview illustration of proposed network architecture of the generator and discriminator

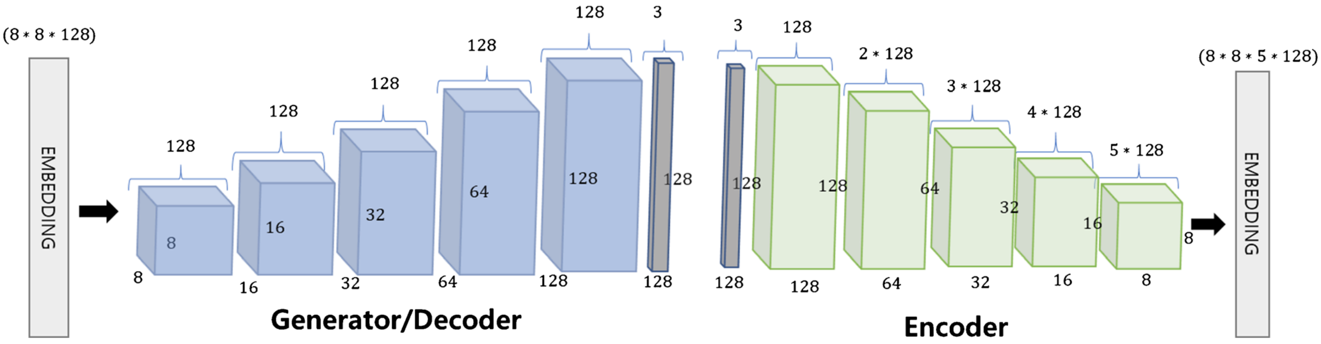

Fig. 2 Proposed network architecture for the generator and discriminator

based on DeLong et al. [21]. All statistics were done using Medcalc and the Analyse-it software program.

Results

Training conditional boundary equilibrium GAN

The network was implemented and trained using a TensorFlow framework. We trained the network using an Adam optimizer with initial learning rate of 10^{-5} and decay factor of 0.95 for each epoch. We set a batch size of 16, and training epochs of 200, and a diversity ratio parameter of 0.7. The diversity ratio parameter was related to the variability of the fake image synthesized by the generator.

In addition, we confirmed the degree of convergence while training. We used the convergence measure (M_{global}) proposed by Berthelot et al. [19]. We measured the M_{global} convergence value at each training iteration of the proposed network. We observed that the generated images became similar to the actual training data as the M_{global} value decreased. The synthetic data generated by the generator during the training process are shown in Fig. 3. The final ADNI data and the synthesized coronal data are shown in Fig. 4.

Performance evaluation of the classification

In the process of training the proposed network, the discriminator was trained to extract appropriate features to distinguish AD patients from NC subjects. In addition, we trained a classifier that consisted of a fully connected network and an SVM based on the features extracted from the discriminator. We compared accuracy, sensitivity, and specificity to measure the performance.

The single-slice performance is illustrated in Table 2 (a). The highest performance was for 48 slice with 92.20%, 89.04%, and 95.59% for the Severance dataset, and 92.23%, 90.13%, and 94.27% for the ADNI dataset, for accuracy, sensitivity, and specificity, respectively. The AUC values were 0.96 and 0.97 for the Severance and ADNI datasets. The double-slice performance is shown in Table 2 (b). In a slice 45,48 pair, classification accuracy, sensitivity, and specificity were 94.33%, 91.78%, and 97.06%, respectively, for Severance data, and 94.82%, 92.11%, and 97.45% for the ADNI dataset. The *p* values for single-slice performance comparison between Severance hospital and ADNI dataset showed no significant statistical difference for the range from 30 to 69 slices. The p values for double-slice model performance comparison between the Severance hospital and ADNI datasets did not show a statistical difference from 30 to 69 slices.

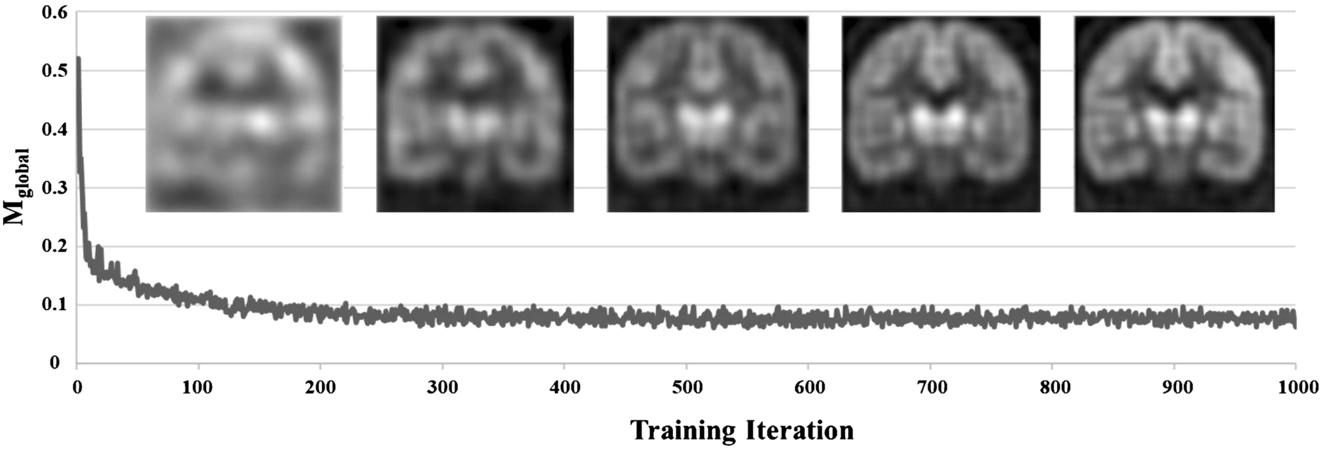

Fig. 3 Synthesized image and convergence parameter according to training iteration of proposed network

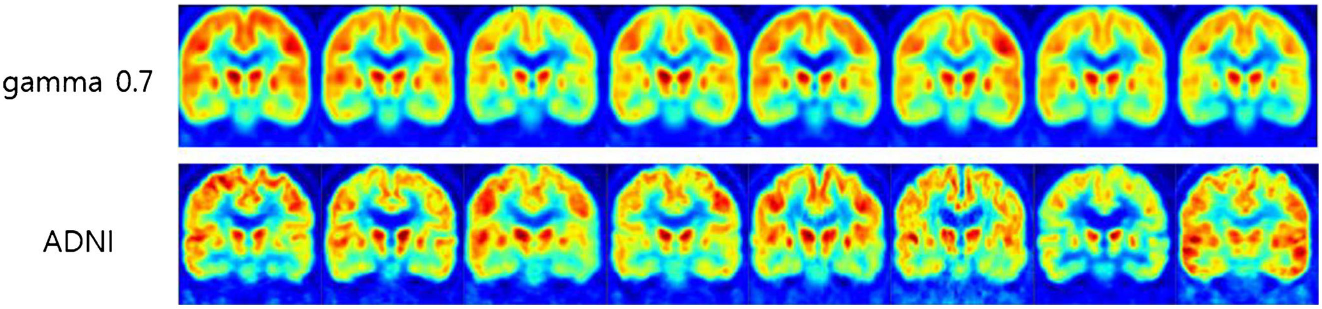

Fig. 4 Fake coronal data synthesized by the generator according to gamma parameter and the ADNI data used to train the generative adversarial network

We used the AUC diagnostic test to compare the performance using single-slice and double-slice inputs. The best performance was shown with double slice of 45 and 48 in both datasets. Using double slice of 45 and 48 as input showed the highest AUC value (0.98) for both datasets. To verify the performance, we used the ROC diagnostic test to determine if there was a statistically significant difference using single-slice or double-slice inputs. Table 3 refers to the ROC

Table 2 Alzheimer's disease classification performance **(a)** using a single slice and **(b)** double slices

Slice number	ACC %		SEN %		SPE %		AUC		*p* value
	Sev	ADNI	Sev	ADNI	Sev	ADNI	Sev	ADNI	
(a)									
30	82.98	85.11	80.82	88.82	85.29	81.53	0.91	0.91	0.42
33	84.40	87.38	78.08	85.53	91.18	89.17	0.90	0.94	0.57
36	85.82	86.73	82.19	83.55	89.71	89.81	0.90	0.95	0.81
39	87.23	88.67	80.82	84.21	94.12	92.99	0.93	0.95	0.95
42	87.23	84.79	86.30	94.08	88.24	75.80	0.92	0.94	0.10
45	88.65	91.59	87.67	90.13	89.71	92.99	0.95	0.97	0.63
48	92.20	92.23	89.04	90.13	95.59	94.27	0.96	0.97	0.85
51	90.07	90.61	91.78	92.11	88.24	89.17	0.95	0.95	0.64
54	75.89	73.46	57.53	58.55	95.59	87.90	0.93	0.84	0.53
57	85.82	86.73	82.19	84.21	89.71	89.17	0.94	0.94	0.91
60	70.21	76.05	43.84	55.92	98.53	95.54	0.92	0.85	0.16
63	85.11	85.44	87.67	92.76	82.35	78.34	0.93	0.92	0.59
66	82.27	83.17	68.49	76.97	97.06	89.17	0.94	0.92	0.20
69	78.01	85.44	76.71	88.82	79.41	82.17	0.86	0.92	0.54
72	73.05	75.40	82.19	51.97	63.24	98.09	0.81	0.91	$p < 0.0001$
(b)									
30, 33	82.98	89.97	80.82	92.11	85.29	87.90	0.89	0.96	0.62
33, 36	84.40	91.59	84.93	96.71	83.82	86.62	0.89	0.98	0.61
36, 39	86.52	89.00	75.34	80.92	98.53	96.82	0.91	0.98	0.06
39, 42	86.52	89.64	83.56	94.74	89.71	84.71	0.91	0.97	0.23
42, 45	88.65	91.26	87.67	96.05	89.71	86.62	0.94	0.97	0.47
45, 48	94.33	94.82	91.78	92.11	97.06	97.45	0.98	0.98	0.65
48, 51	93.62	94.17	90.41	93.42	97.06	94.90	0.96	0.98	0.95
51, 54	92.20	89.00	89.04	84.21	95.59	93.63	0.96	0.95	0.48
54, 57	87.23	85.44	84.93	79.61	89.71	91.08	0.95	0.92	0.30
57, 60	87.23	84.79	90.41	87.50	83.82	82.17	0.94	0.92	0.62
60, 63	85.11	86.41	82.19	86.18	88.24	86.62	0.91	0.93	0.85
63, 66	86.52	86.41	78.08	86.18	95.59	86.62	0.93	0.92	0.19
66, 69	84.40	84.47	78.08	84.87	91.18	84.08	0.90	0.92	0.31
69, 72	80.85	82.20	84.93	67.11	76.47	96.82	0.88	0.93	$p < 0.0001$

ACC accuracy, *SEN* sensitivity, *SPE* specificity, *AUC* area under the curve, *Sev* Severance hospital dataset

Table 3 ROC diagnostic test with single slice and double slice using ADNI dataset and Severance dataset

Dataset	Slice number		Difference between areas (Standard error)	95% Confidence interval	p value
ADNI	42, 45	42	0.0350 (0.0093)	0.0167 to 0.0532	0.0002 ($p < 0.05$)
		45	0.0073 (0.0075)	−0.00732 to 0.0220	0.3259
		48	0.0022 (0.0084)	−0.0143 to 0.0188	0.7868
		51	0.0237 (0.0101)	0.00400 to 0.0434	0.0184 ($p < 0.05$)
	45, 48	42	0.0433 (0.0114)	0.0210 to 0.0657	0.0001($p < 0.05$)
		45	0.0157 (0.0077)	0.000675 to 0.0308	0.0406 ($p < 0.05$)
		48	0.0107 (0.0053)	0.000337 to 0.0210	0.043 ($p < 0.05$)
		51	0.0321 (0.0106)	0.0113 to 0.0529	0.0025 ($p < 0.05$)
Severance	42,45	42	0.015 (0.0138)	−0.0192 to 0.0504	0.0281
		45	0.0156 (0.0177)	− 0.0164 to 0.0633	0.3790
		48	0.0235 (0.0203)	−0.0395 to 0.0538	0.2481
		51	0.0071 (0.0238)	0.0188 to 0.0944	0.7638
	45,48	42	0.0566 (0.193)	0.0188 to 0.0944	0.0033 ($p < 0.05$)
		45	0.026 (0.0112)	0.00396 to 0.0480	0.0208 ($p < 0.05$)
		48	0.0181 (0.009)	0.00138 to 0.0349	0.0338 ($p < 0.05$)
		51	0.0344 (0.0167)	0.00173 to 0.0672	0.0391 ($p < 0.05$)

diagnostic test results for the ADNI and Severance hospital datasets, respectively. We found that there were statistical differences using double slice 45,48 when compared to adjacent slices 42 to 51 in both datasets, whereas double slice 42,45 showed no statistical difference in single slice in both datasets. Supplementary Fig.1 illustrate the ROC curve of comparing double slice 45,48 for the ADNI and Severance hospital datasets.

To compare the proposed model with other AD classification models, the model in Glozman et al. [7] was used. Table 4 shows the performance of the proposed model and other AD classification models. The p value was calculated using the chi-square test to compare the generalization ability of the other AD classification models. The other AD classification models listed in Table 4 did not show generalization ability. The performance of the ADNI and Severance datasets significantly differed ($p < 0.05$). As shown in Table 5, ROC statistical methods were used to show those differences in performance. The results demonstrated that the proposed method was significantly higher than other AD classification methods in performance. Moreover, similar to the findings with our method, double slice showed better performance than single slice in other AD classification methods. Therefore, the proposed model is not only generalized in the Severance model but also has high performance when compared to other AD classification models. Supplementary Fig.2 illustrates the ROC curve of comparison of our methods and other AD classification models.

To compare performance depending on the number of input slices, we used single slices, quadruple slices, and 3D volumes as input data. The best performance of single slice is in Table 2 (a). Quadruple slice or 3D volume data was selected around slice 45,48. The ROC diagnostic test statistical method was used to see if there were significant differences in number of input slices. Table 6 shows the result depending on the number of inputs. In the Severance data, the double slice model showed the best performance with significant differences with other models. In the ADNI data, 3D volume data showed the highest performance, but there were no significant differences compared to the double slice model. We considered this result as model overfitting to the

Table 4 Alzheimer's disease classification performance using our methods and other AD classification models

Model	Slice number	ACC %		SEN %		SPE %		AUC		p value
		Sev	ADNI	Sev	ADNI	Sev	ADNI	Sev	ADNI	
Our method	45,48	94.33	94.82	92.11	91.78	97.45	97.06	0.98	0.98	0.623
Glozman et al. [7]	45	74.47	75.73	95.59	67.76	54.79	83.44	0.884	0.844	0.018
	48	73.76	79.61	56.16	77.63	92.65	81.53	0.874	0.854	0.001
	45,48	81.56	83.5	85.29	75.8	78.08	91.45	0.898	0.927	0.013

ACC accuracy, *SEN* sensitivity, *SPE* specificity, *AUC* area under the curve, *Sev* Severance hospital dataset

Table 5 ROC diagnostic test with our method and other AD classification model

Dataset	vs Our method	Difference between areas (Standard error)	95% Confidence interval	p value
Severance	Slice 45	0.0965 (0.0306)	0.0365 to 0.156	0.0016
	Slice 48	0.107 (0.0306)	0.0468 to 0.167	0.0005
	Slice 45,48	0.0824 (0.0273)	0.0289 to 0.136	0.0025
ADNI	Slice 45	0.139 (0.0225)	0.0946 to 0.183	$p < 0.0001$
	Slice 48	0.129 (0.0218)	0.0860 to 0.171	$p < 0.0001$
	Slice 45,48	0.0556 (0.0166)	0.0232 to 0.0881	0.0008

ADNI dataset. The difference in properties of the datasets led to lower performance in the Severance dataset.

Discussion

In this study, we proposed a GAN-based method to classify AD and NC from [^{18}F] FDG PET/CT images for external validation based on a training dataset using the ADNI open database. Because there were no previous studies regarding external validation for AD classification, we used our model to learn generalized features from preprocessed ADNI data. Our model creation was a two-stage process, training GAN using slice selective learning for feature extraction and an SVM classifier for AD classification. The method applied is practical when using a limited number of [^{18}F] FDG PET/CT images for training a deep learning model.

Generally, deep learning models require a large amount of data for training. However, it is difficult to collect such large amounts of [^{18}F] FDG PET/CT images taken for diagnosis of AD and NC at individual hospitals. In addition, it is widely known that privacy, healthcare industry standards, and incomplete integration of medical information systems (and related issues) make medical datasets difficult to collect compared to general image data. More importantly, [^{18}F] FDG PET/CT images show different properties depending on various acquisition instruments, collection environment, acquisition protocols, and reconstruction methods, which hamper use of the network [22]. Therefore, the model performance of external validation shows lower performance compared to the training dataset.

To overcome these challenges, we designed our model with enhanced regularized performance. We set a range that covered the most important AD-related regions and searched for the most appropriate slices for classification. Two-dimensional sectional images extracted from 3D PET volumes were used. Applying slice-selective learning reduced the number of training parameters and helped for training our model with unbiased features. To enhance the generalization performance, we evaluated performance with added slices. Table 3 indicates that using double-slice input statistically improved the performance more than with single-slice input, when an appropriate slice was chosen. Table 6 demonstrates that double slice is the optimal number of slices compared to other numbers of input slices. Moreover, we observed that double slice shows better performance than single slice when applied to other AD classification methods. Table 4 illustrates the result.

Table 2 shows the results of the range from the amygdala to the posterior end of the PCC. We noticed that the imaging acquisition environment may hamper the performance on the ADNI and our independent datasets. This is because the partial volume effect (PVE) may affect the small regions to blur the region of hippocampus (HIP) and entorhinal cortex (EC) [11]. However, PVE occurs due to limited resolution of the PET imaging system. This implies that the imaging acquisition environment, such as the imaging acquisition

Table 6 ROC diagnostic test with number of input slice using single, double, quadruple, and 3D volume

Number of input slices	Severance					ADNI				
	ACC	SEN	SPE	AUC	p value	ACC	SEN	SPE	AUC	p value
Double(our method)	94.33	92.11	97.45	0.98	__	94.82	91.78	97.06	0.98	__
Single	92.20	89.04	95.59	0.96	0.03	92.23	90.13	94.27	0.97	0.04
Quadruple	84.39	91.78	76.47	0.93	0.01	93.85	96.71	91.08	0.98	0.46
3D volume	85.1	90.41	79.41	0.94	0.01	95.14	96.05	94.26	0.98	0.82

ACC accuracy, *SEN* sensitivity, *SPE* specificity, *AUC* area under the curve instrument, hindered performance on datasets. Therefore, the slice 45 and 48 showed the best performance, which covers the PCC region. These ranges are less affected by the PVE, compared to HIP and EC [11].

Although we developed a model with enhanced generalization performance, there were limitations. First, we only performed a feasibility test with one external validation dataset. We need to test using more independent datasets in further research. Second, we relied on coronal planes for this study because neuropathologic changes of autopsy specimens are assessed using coronal slices. Examination using axial and sagittal planes is also needed. Third, the performance comparison should be performed with variations from 5 mm to 7.5 mm slice intervals.

Conclusions

We used the proposed GAN-based model with external validation to classify AD and NC from among selected [^{18}F] FDG PET/CT images associated with AD neuropathologic changes. The model showed diagnostic performance for an ADNI dataset that was similar to that with a different dataset from our hospital. When there are insufficient datasets to train individual deep neural networks with single-source training datasets, our approach seems a feasible alternative to classify AD and NC using datasets from various hospitals.

Acknowledgments We thank Severance Hospital of the Department of Nuclear Medicine for their help with this project.

Funding information This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2018M3C7A1056898), and in part by the Bio & Medical Technology Development Program of theNational Research Foundation (NRF) funded by theMinistry of Science and ICT. (NRF-2018M3A9H6081483)

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflicts of interest.

Ethical approval All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants in the study.

References

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007;3:186–91.
2. Mosconi L, Berti V, Glodzik L, Pupi A, De Santi S, De Leon MJ. Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. J Alzheimers Dis. 2010;20:843–54.
3. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48 Available from: http://www.ncbi.nlm.nih.gov/pubmed/28301734 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5479722.
4. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: ImageNet Classification with Deep Convolutional Neural Networks. Curran Associates, Inc.; p. 1097–105 2012. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-w.
5. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng. 2015;62:1132–40.
6. Suk HI, Shen D. Deep learning-based feature representation for AD/MCI classification. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 8150 LNCS: 583–90, 2013.
7. Glozman T, Liba O. Hidden cues: deep learning for Alzheimer's disease classification CS331B project final report. 2016.
8. Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based multi-class classification of Alzheimer's disease using MRI. IST 2017 - IEEE Int Conf Imaging Syst Tech Proc. 2018-Janua:1–6, 2018.
9. Hosseini-asl E, Keynton R, El-baz A, Drzezga A, Lautenschlager N, Siebner H, et al. Alzheimer ' s disease diagnostics by adaptation of 3D convolutional network electrical and computer engineering department , University of Louisville , Louisville , KY, USA. Eur J Nucl Med Mol Imaging. 2016;30:1104–13.
10. Liu M, Cheng D, Wang K, Wang Y. Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics. 2018;16:295–308.
11. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007;48:932–45.
12. Jagust WJ, Landau SM, Koeppe RA, Reiman EM, Chen K, Mathis CA, et al. The Alzheimer's Disease Neuroimaging Initiative 2 PET Core. Alzheimer's Dement. 2015:2015.
13. Ashburner J, Friston KJ. Nonlinear spatial normalization using basis functions. Hum Brain Mapp. 1999.
14. Mosconi L, Mistur R, Switalski R, Tsui WH, Glodzik L, Li Y, et al. FDG-PET changes in brain glucose metabolism from normal cognition to pathologically verified Alzheimer's disease. Eur J Nucl Med Mol Imaging. 2009;36:811–22.
15. De Santi S, De Leon MJ, Rusinek H, Convit A, Tarshish CY, Roche A, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol Aging. 2001;22:529–39.
16. Drzezga A, Lautenschlager N, Siebner H, Riemenschneider M, Willoch F, Minoshima S, et al. Cerebral metabolic changes accompanying conversion of mild cognitive impairment into Alzheimer's disease: a PET follow-up study. Eur J Nucl Med Mol Imaging. 2003.
17. Powers JM. Practice guidelines for autopsy pathology. Autopsy procedures for brain, spinal cord, and neuromuscular system. Autopsy Committee of the College of American pathologists. Arch Pathol Lab Med. 119:777, 1995–83.
18. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks 1–9, 2014. Available from: http://arxiv.org/abs/1406.2661.

19. Berthelot D, Schumm T, Metz L. BEGAN: Boundary Equilibrium Generative Adversarial Networks. arXiv. 2017.
20. Mirza M, Osindero S. Conditional Generative Adversarial Nets 1–7, 2014. Available from: http://arxiv.org/abs/1411.1784.
21. 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 Available from: http://www.ncbi.nlm.nih.gov/pubmed/3203132.
22. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50:11S–20S.

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