PLOS ONE

RESEARCH ARTICLE

Performance evaluation in [^{18}F]Florbetaben brain PET images classification using 3D Convolutional Neural Network

Seung-Yeon Lee[1,2], **Hyeon Kang**[2], **Jong-Hun Jeong**[3], **Do-young Kang**[1,2,4]*

1 Department of Translational Biomedical Sciences, Dong-A University, Busan, Korea, **2** Institute of Convergence Bio-Health, Dong-A University, Busan, Korea, **3** DEEPNOID Inc., Seoul, Korea, **4** Department of Nuclear Medicine, Dong-A University Medical Center, Busan, Korea

* dykang@dau.ac.kr

Citation: Lee S-Y, Kang H, Jeong J-H, Kang D-y (2021) Performance evaluation in [^{18}F]Florbetaben brain PET images classification using 3D Convolutional Neural Network. PLoS ONE 16(10): e0258214. https://doi.org/10.1371/journal.pone.0258214

Editor: Pierpaolo Alongi, Fondazione Istituto G. Giglio di Cefalu, ITALY

Received: April 8, 2021

Accepted: September 21, 2021

Published: October 20, 2021

Data Availability Statement: All relevant data are available on Figshare: https://doi.org/10.6084/m9.figshare.16702519.v1.

Funding: This work was supported by the Dong-A University research fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

High accuracy has been reported in deep learning classification for amyloid brain scans, an important factor in Alzheimer's disease diagnosis. However, the possibility of overfitting should be considered, as this model is fitted with sample data. Therefore, we created and evaluated an [^{18}F]Florbetaben amyloid brain positron emission tomography (PET) scan classification model with a Dong-A University Hospital (DAUH) dataset based on a convolutional neural network (CNN), and performed external validation with the Alzheimer's Disease Neuroimaging Initiative dataset. Spatial normalization, count normalization, and skull stripping preprocessing were performed on the DAUH and external datasets. However, smoothing was only performed on the external dataset. Three types of models were used, depending on their structure: Inception3D, ResNet3D, and VGG3D. After training with 80% of the DAUH dataset, an appropriate model was selected, and the rest of the DAUH dataset was used for model evaluation. The generalization potential of the selected model was then validated using the external dataset. The accuracy of the model evaluation for Inception3D, ResNet3D, and VGG3D was 95.4%, 92.0%, and 97.7%, and the accuracy of the external validation was 76.7%, 67.1%, and 85.3%, respectively. Inception3D and ResNet3D were retrained with the external dataset; then, the area under the curve was compared to determine the binary classification performance with a significance level of less than 0.05. When external validation was performed again after fine tuning, the performance improved to 15.3%p for Inception3D and 16.9%p for ResNet3D. In [^{18}F]Florbetaben amyloid brain PET scan classification using CNN, the generalization potential can be seen through external validation. When there is a significant difference between the model classification performance and the external validation, changing the model structure or fine tuning the model can help improve the classification performance, and the optimal model can also be found by collaborating through a web-based open platform.

1 Introduction

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disease in which a patient gradually loses memory, mental function, and the ability to continue daily activities [1]. As there is no effective treatment for AD, an accurate diagnosis is essential for developing the patient's future treatment plan. Neuroimaging technology using AD-related biomarkers is widely used to increase the reliability of AD diagnosis [2]. Radiotracers are among the types of biomarkers that can be injected into the subject's body and observed via PET. Representative radiotracers used to diagnose AD include 2-[^{18}F]fluoro-D-glucose, which can identify the degree of brain metabolism, and [^{18}F]Florbetaben, [^{18}F]Florbetapir and [^{18}F]Flutemetamol which can observe brain amyloid plaque load.

According to specific or diagnostic criteria [3–5], several systems apply brain PET scans to machine learning and deep learning models to train, evaluate, and classify images. Moreover, there is also recent work using them on automatic and semi-automatic segmentation algorithms in PET [6, 7]. In one study, an AD diagnosis classifier using PCA and SVM was utilized after image dimension reduction of [^{18}F]Florbetaben brain PET images [8], and in another, images were classified according to amyloid deposition with an accuracy of 89% [9] using Visual Geometry Group (VGG) 16 [10], which is a well-known structure among convolutional neural networks (CNN) [11, 12] that specializes in image feature extraction using deep learning technology.

Meanwhile, because of the nature of medical images, acquisition costs can be high; thus, it is not easy to construct a large dataset. When a model is trained and tested with a limited number of datasets, to confirm the generalization possibility, it is necessary to configure an external dataset that is different from the source of the training dataset and to validate its potential.

Previous studies, such as the external validation of pancreatic cancer from CT images [13] and the external validation of malignancy risk prediction of lung nodules [14], have been reported. It is necessary to apply the validation process to brain imaging to build a model and then to externally validate the model with the same type of brain images obtained from other sites.

In this study, brain PET scans of [^{18}F]Florbetaben, a diagnostic radiotracer that visualizes the classification of β-amyloid (Aβ), the main component of amyloid plaques found in the brain, were acquired from Dong-A University Hospital (DAUH) and used to construct a dataset. We created models that receive 3D voxel input by deriving characteristic structures from the well-known CNN structures Inception, VGG, and ResNet. We then selected one representative model for each structure after training according to the model selection criteria. With this model, the images acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI) were used as a dataset and externally validated to examine the possibility of generalization.

2 Materials and methods

2.1 Data acquisition

The DAUH dataset, listed in Table 1, comprises a total of 432 subjects according to the visually assessed criteria of [^{18}F]Florbetaben PET. An available database was collected from 2015 to 2020 from the Department of Nuclear Medicine at DAUH for a possible population of [^{18}F]Florbetaben PET. Labeling for Aβ negative and positive was performed according to the decision of a nuclear medicine specialist at DAUH.

PET scans of the possible population were acquired using a Biograph 40 mCT Flow PET/CT Scanner (Siemens Healthcare, Knoxville, TN, USA) and reconstructed via UltraHD-PET

Table 1. Demographics and positivity of study participants according to each dataset group.

Dataset	DAUH		External	
n	432		251	
Visual Assessment	Negative	Positive	Negative	Positive
Subjects	191	241	142	109
Age	68.1	69.9	70.8	73.7
Sex, male %	35.6	44.4	40.8	56.9
Education	9.1	10	16.1	16

https://doi.org/10.1371/journal.pone.0258214.t001

(TrueX-TOF) to obtain PET images, and the obtained images were processed with a 3mm FWHM Gaussian filter. All patients underwent a 20 min positron emission scan at 90 min after intravenous injection of 300mBq of [^{18}F]Florbetaben (NeuraCeq, Piramal, Mumbai, India), and the helical CT scans were acquired with a 0.5 s rotation time at 100 kVp and 228 mAs. The images were finally stored in the DICOM format.

Finally, [^{18}F]Florbetaben PET images collected from the Alzheimer's Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI) were used as an external dataset in this study. The database includes scans of subjects with normal control, mild cognitive impairment (MCI), and AD. These scans were preprocessed as an ADNI internal protocol and co-registration, averaging, size changing, standardization, and smoothing processes were performed. However, since ADNI does not determine whether the [^{18}F]Florbetaben scans are Aβ negative or positive, classification was performed by the nuclear medicine specialists at DAUH.

Unlike the DAUH dataset, images from the external dataset were not processed with 3 mm FWHM Gaussian filters. To achieve the same conditions, a 3 mm FWHM Gaussian filter was used to process the statistical parametric mapping (SPM) after the acquisition. An example of Gaussian filter processing is shown in Fig 1.

2.2 Image preprocessing in common

We performed spatial normalization and count normalization to classify Aβ deposition using brain PET. We used the SPM library [15] based on MATLAB. All dataset groups applied the same preprocessing procedure using the same protocol, even if there were differences in the source and collection times.

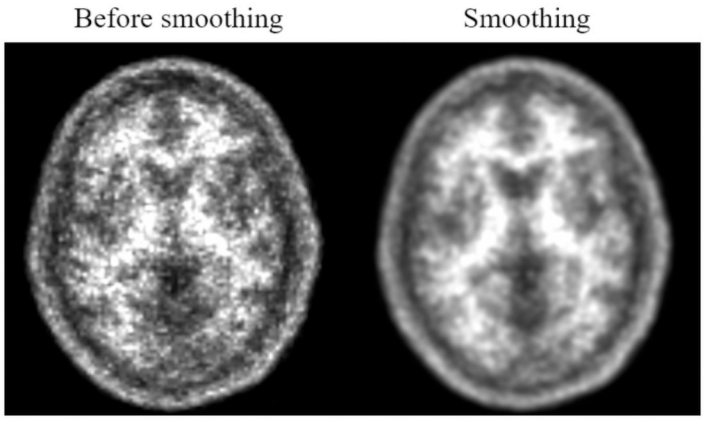

Fig 1. Smoothing of the external data. After the smoothing process, an image in which noise is reduced can be obtained.

https://doi.org/10.1371/journal.pone.0258214.g001

Spatial normalization is the registration of the original image to a specific PET template. Subjects with various brain shapes can be compared through image registration by mapping the subject's PET image to a reference brain template [16].

Meanwhile, we made a reference brain PET template in-house. Co-registration was performed with a PET image and a paired CT image obtained from a subject. The template images are average images of 20 normal brains, 20 Alzheimer's dementia brains, and 40 left and right inversion images, for a total of 80 images. After image registration, the tissue is stretched or compressed to fit the template brain. We can average the PETs of multiple individuals to reconstruct them into a reference brain space and provide atlas anatomical addresses mapped to the same reference brain space at the data locations in the image. We also performed a cropping process that cuts the empty space around the brain, reducing its size from $400 \times 400 \times 110$ to $95 \times 79 \times 68$.

Count normalization is intended for numerical comparisons between images because the image intensity level varies due to differences in the number of radioactive isotopes administered, individual characteristics, or individual body conditions. This normalization was performed assuming that the absorption of radioactive tracers in a brain region is constant for each person. Count normalization normalizes the entire observation area to the value of the area representing non-specific, lesion-independent absorption, allowing absolute and relative comparisons in specific absorption areas of the patient-patient image [17].

The [^{18}F]Florbetaben radiotracer exhibits non-specific uptake in the cerebellar region and specific uptake in the gray matter region of the cerebrum; thus, count normalization was performed by using the cerebellar region of the PET template applied in spatial normalization to each patient image.

In addition, the skull, a non-brain tissue, is included in the image because it is spatially normalized with a CT-driven PET template. The presence of these non-brain tissues is considered an obstacle in brain image analysis. Therefore, in brain imaging analysis studies, a preprocessing commonly referred to as skull stripping is required [18].

In the [^{18}F]Florbetaben amyloid brain PET classification model, spatial normalization, count normalization, and skull stripping are commonly performed for all datasets, and an example of a brain PET scan image that has been preprocessed is shown in Fig 2.

2.3 Model architecture

In this study, we apply CNNs of three well-known architectures to the amyloid classification problem in [^{18}F]Florbetaben brain PET. The architectures considered are Inception [19, 20], ResNet [21], and VGG19 [10]. The reasons for choosing these models were to achieve the best performance in various tasks, use a small kernel (3×3) [10] instead of a large kernel, utilize a deep but sparse network structure [19], and provide residual connectivity [21]. We implemented the model ourselves by adopting parts of the primary characteristics of these structures and applying 3D convolution filter. Each representative structure is shown in Fig 3.

To briefly summarize the meaning of each layer:

- Conv3D: 3D convolution kernel layer.
- BatchNormalization: To avoid gradient vanishing problems due to structural complexity and speed up the training [22].
- Activation: When the data that has passed through the layer is transmitted to the next layer, it plays a role in determining whether to transmit the input data according to a specific criterion, and all model structures in this study apply the Rectifier Linear Unit (ReLU) [23] activation function.

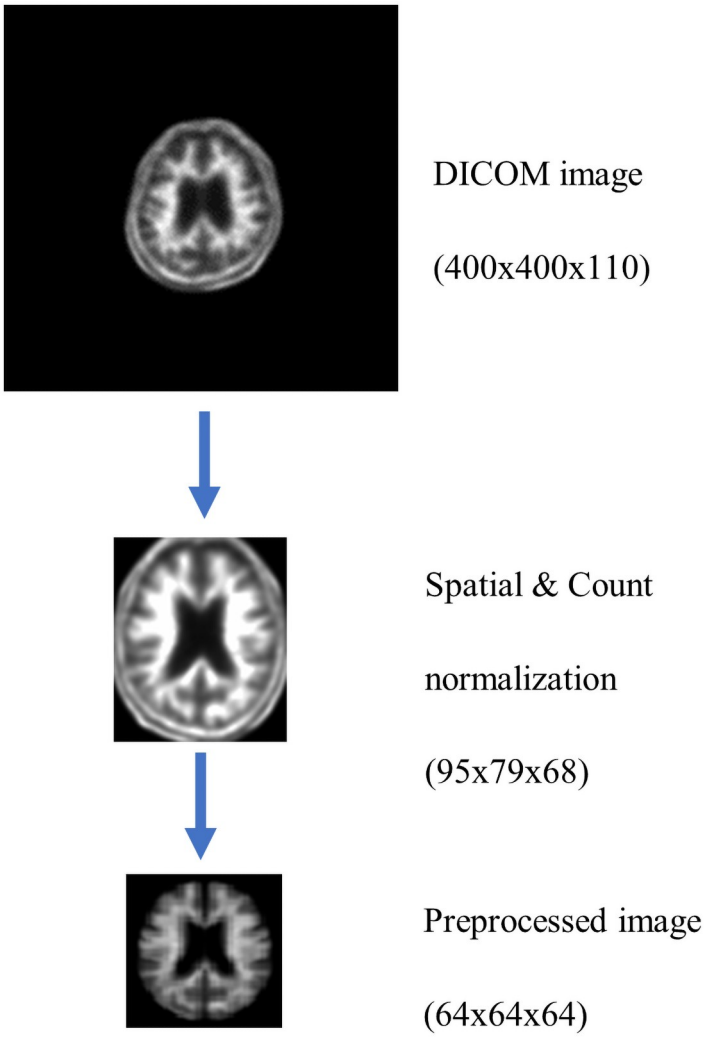

Fig 2. Commonly applied preprocessing.

https://doi.org/10.1371/journal.pone.0258214.g002

- Concatenate: To connect data horizontally.
- Pooling: Also known as sub-sampling, the image is reduced by selecting a large value or taking an average in the corresponding receptive field.
- Dropout: Only certain weights are kept at a certain probability and the remaining connected units are diluted. This is known to prevent network overfitting [24].

Recently, 3D CNN research has been actively conducted to extract features of 3D medical images for classification [25]. Moreover, because brain images are volume data, 3D CNNs can be configured for image classification to extract 3D spatial features from 3D PET images. Accordingly, Inception, ResNet, and VGG models are constructed in the form of a 3D CNN and are named Inception3D, ResNet3D, and VGG3D in our experiment.

2.4 Model selection and evaluation

Eighty percent of the DAUH dataset, as shown in Fig 4, was used for model training and the rest is used for model evaluation. The samples were splitted with Stratified ShuffleSplit cross-

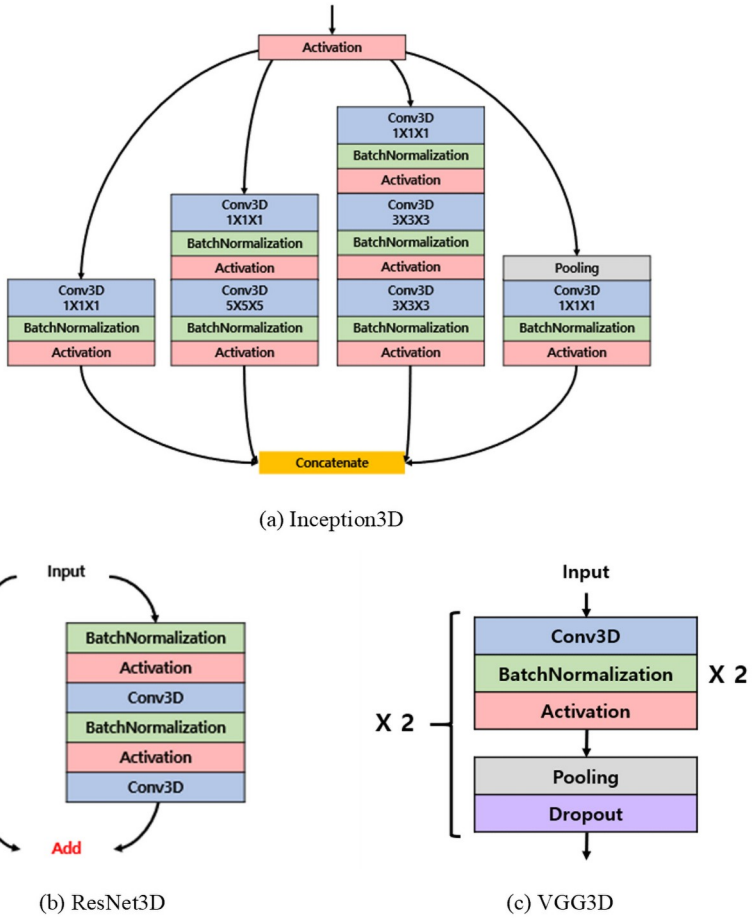

Fig 3. Primary structure of the models. (a) and (b) are the most characteristic parts of the structure, and (c) is the overall structure of the model. The '×2' written on the side means that the process is repeated as many times as the number.

https://doi.org/10.1371/journal.pone.0258214.g003

Dong-A University Hospital(DAUH) dataset

Years: 2015-2020 Total: 432

Amyloid	subjects
Positive	241
Negative	191

For model training

Amyloid	Subjects
Positive	192
Negative	153

Model evaluation set

Amyloid	Subjects
Positive	49
Negative	38

External dataset from ADNI

Total: 251

Amyloid	subjects
Positive	109
Negative	142

External validation set

Amyloid	Subjects
Positive	71
Negative	92

Fig 4. DAUH and external datasets for model training and validation.

https://doi.org/10.1371/journal.pone.0258214.g004

validator (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html), which was randomly extracted from each group labeled amyloid positive and negative. Along with the training set, there was also a validation set used to ensure that the model is trained well. This is explained in more detail in Fig 5.

The model training was conducted by four-fold cross-validation with 192 amyloid positive subjects and 153 amyloid negative subjects from the DAUH dataset. If the dataset is small, the reliability of the performance evaluation is reduced. If the performance varies depending on how the validation set is held, the effect of the match will bias the model evaluation performance. To solve this, cross-validation ensures that all data are used as a validation set at least once.

As shown in Fig 6, the entire dataset can be divided into four subsets; the first subset is used as the validation set in the first iteration, and the remaining subsets are used as the training set. In the second iteration, the second subset is used as the validation set, and the remaining subsets are used as the training set. By repeating the number of subsets in this manner, we could select the lowest loss model out of the four performances.

All models of each structure were selected through four-fold cross-validation [26] and evaluated with the model evaluation set of the DAUH dataset specified in Fig 4.

In the external dataset, some samples were randomly extracted using the same sample extraction method. If the accuracy was significantly different between the model evaluation and the external validation, it is possible that a part of the external dataset could be included in the training set to acquire generalization performance. In that case, only a portion of the dataset was composed of the external validation set.

3 Results

We investigated information about the model and the time required for each experiment. Specifications about the models can be viewed in Table 2. Experiment time consists of training, image loading, and prediction using the validation set. The required time for each network is Inception3D 1.46hrs, ResNet3D 1.9hours, and VGG3D 1.90 hours. Inference time took about 0.24 seconds per subject, and all experiments were performed on a workstation with four NVIDIA Titan Xp GPUs.

3.1 Data distribution

To confirm the data distribution of the preprocessed DAUH and external datasets before CNN-based analysis, as shown in Fig 7, t-SNE (Stochastic Neighbor Embedding) [27] visualization, which is widely used for visualization after a data dimension reduction, was performed. Although the sources of the datasets are different, the same groups have a similar distribution. In other words, data similarity was observed between the same groups.

3.2 Model evaluation

The model evaluation results of the Inception3D, ResNet3D, and VGG3D models after training the models with the DAUH training dataset are summarized in Table 3. To evaluate the performance of each model, three metrics were considered: sensitivity, specificity, and accuracy.

Sensitivity is the proportion of subjects who are inferred to be positive among all Aβ-positive subjects and is defined as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

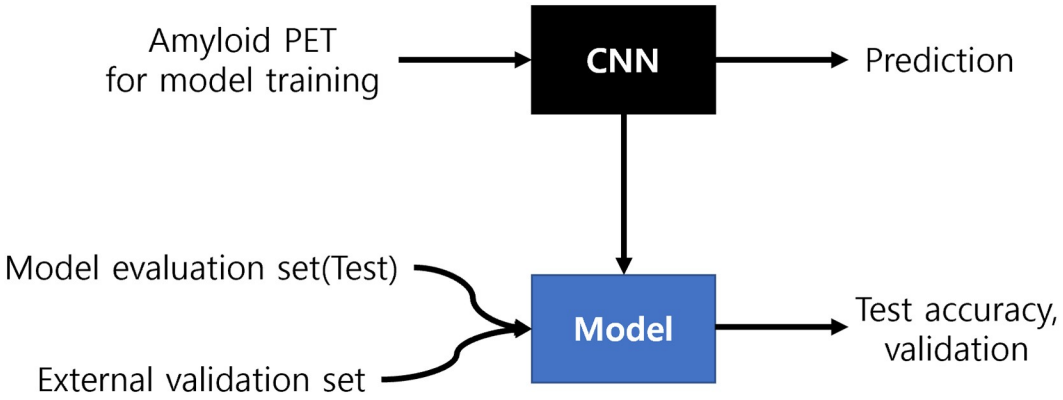

Fig 5. Overall workflow.

https://doi.org/10.1371/journal.pone.0258214.g005

where TP represents the number of true positives and FN represents the number of false negatives.

Specificity is the proportion of subjects who are inferred to be negative among all Aβ-negative subjects and is defined as follows:

$$Specificity = \frac{TN}{TN + FP}$$

where TN represents the number of true negatives and FP indicates the number of false positives.

PPV(Positive Predictive Value) is the probability that those that come out Aβ positive actually have Aβ positive according to the ground truth. NPV(Negative Predictive Value) is the probability that those that come out Aβ negative actually have Aβ negative according to the ground truth.

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{FN + TN}$$

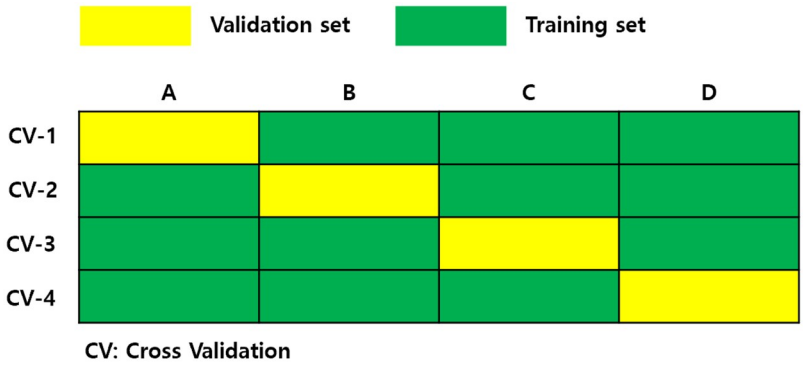

Fig 6. Four-fold cross-validation for each iteration.

https://doi.org/10.1371/journal.pone.0258214.g006

Table 2. Specifications about the models.

	Inception3D	ResNet3D	VGG3D
Total params	6,098,530	17,620,196	17,236,386
Trainable params	6,093,538	17,612,642	17,236,002
Non-trainable params	4,992	7,554	384
Size on disk	70	202	195

https://doi.org/10.1371/journal.pone.0258214.t002

Accuracy is the degree of closeness between the predicted value and the actual value of the subjects. It is calculated as the number of true positives and negatives among all accurate and evaluated subjects as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

We plotted the receiver operating characteristic (ROC) curve of our method for Aβ positivity, as shown in Fig 8, and calculated the area under the curve (AUC). The AUC value was close to 1 for all three models.

Table 4 summarizes the pairwise comparison results of ROC curves obtained from the Inception3D, ResNet3D, and VGG3D inferences that performed model evaluation. At the 95% significance level, since both p-values are greater than 0.05, the corresponding two comparison areas are not significantly different.

3.3 External validation

With respect to accuracy and AUC, VGG3D showed the best classification performance. As summarized in Table 3, the AUC of the external validations was lower than that of the model evaluation, and a comparison of the p-values summarized in Table 4 indicates that Inception3D and ResNet3D were below the significance level; thus, there were significant differences in the performance evaluation. However, there was no significant difference between the model evaluation and the external validation for the VGG model, as shown by the significance level, p = 0.1950. The ROC curves of the comparison are plotted in Fig 9.

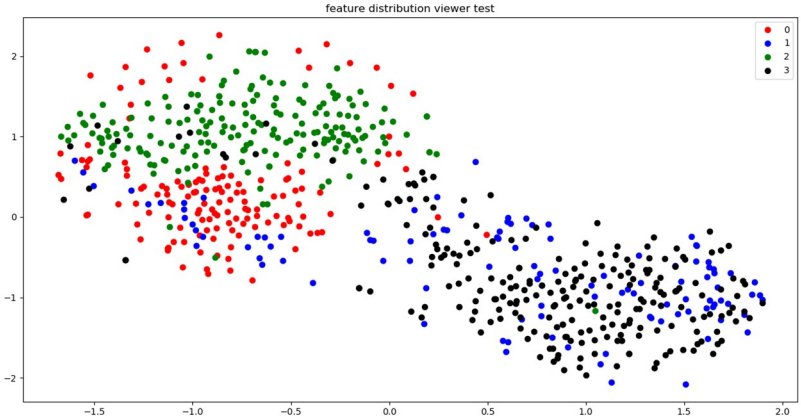

Fig 7. Dataset distribution using t-SNE visualization.

https://doi.org/10.1371/journal.pone.0258214.g007

Table 3. Classification evaluation metrics.

	Inception3D		ResNet3D		VGG3D	
	Model Evaluation	External Validation	Model Evaluation	External Validation	Model Evaluation	External Validation
AUC	0.996	0.883	0.968	0.901	0.98	0.945
Sensitivity	0.918	0.845	0.918	0.944	0.959	0.831
Specificity	1	0.707	0.921	0.469	1	0.945
PPV	1	0.69	0.938	0.458	1	0.831
NPV	0.905	0.855	0.897	0.918	0.95	0.87
Accuracy	0.954	0.767	0.92	0.671	0.977	0.87
Standard Deviation	0.487	0.452	0.482	0.351	0.95	0.853
95% Confidence Interval	0.950 to 1.000	0.823 to 0.928	0.907 to 0.994	0.845 to 0.942	0.923 to 0.998	0.898 to 0.974

https://doi.org/10.1371/journal.pone.0258214.t003

3.4 Retraining model

This is a method of redefining the model to create a generalized model by additionally retraining a part of the external dataset to the model trained with the DAUH training set. In this part, excluding the external validation set shown in Fig 4, the 38 Aβ positives and 50 Aβ negatives were configured as a retraining set for fine tuning. the pre-trained model was imported as it is, and all layers were retrained.

As summarized in Table 5, when the external validation was performed, the classification performance was improved compared to before retraining. The evaluation performance of the models trained with DAUH was compared with that of models additionally retrained with the residue of the external datasets by independent ROC. Both evaluation performances were within the 95% confidence level (Inception3D: 0.5124, ResNet3D: 0.3247), as shown in Table 6. No significant difference was observed between the two evaluations for each model, and the ROC curves are shown in Fig 10.

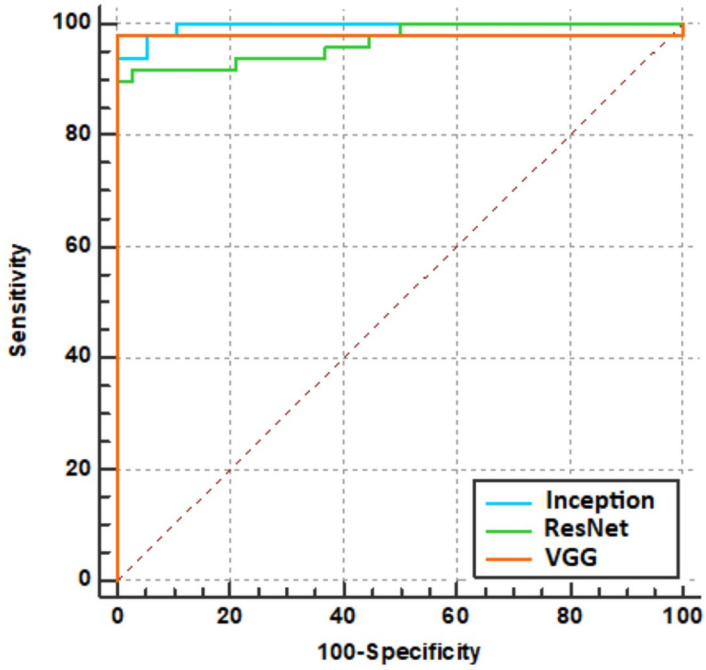

Fig 8. ROC curves for each model evaluation.

https://doi.org/10.1371/journal.pone.0258214.g008

Table 4. Comparison of ROC curves.

Pairwise comparison of ROC curves for each model evaluation				
	Difference between areas	SE	95% CI	P-value
Inception3D vs. ResNet3D	0.0274	0.0154	-0.00282 to 0.0576	0.0756
Inception3D vs. VGG3D	0.0161	0.0208	-0.0246 to 0.0569	0.4384
ResNet3D vs. VGG3D	0.0113	0.0203	-0.0285 to 0.0510	0.5780
Comparison of independent ROC curves with the model evaluations and the external validations				
Inception3D	0.113	0.0266	0.0610 to 0.165	<0.05
ResNet3D	0.0672	0.0306	0.00733 to 0.127	<0.05
VGG3D	0.0349	0.0269	-0.0179 to 0.0876	0.1950

Note. SE = standard error, CI = confidence interval

https://doi.org/10.1371/journal.pone.0258214.t004

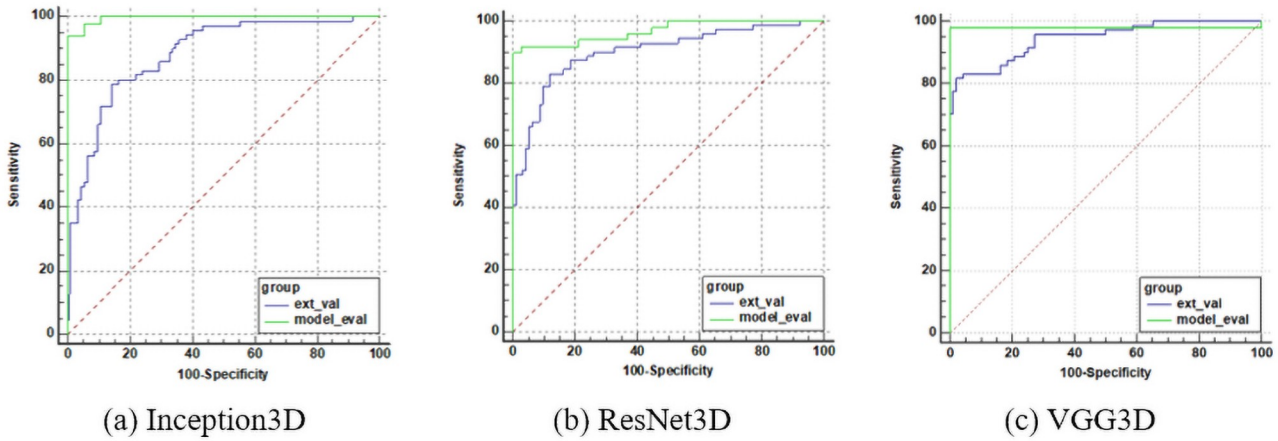

Fig 9. ROC curve comparison for each model. In the legend of each plot, '*ext_val*' means the external validation, and '*model_eval*' means the model evaluation.

https://doi.org/10.1371/journal.pone.0258214.g009

Table 5. External validation after model retraining.

	Inception3D	ResNet3D
AUC	0.95	0.943
Sensitivity	0.845	0.915
Specificity	0.967	0.783
Accuracy	0.92	0.84

https://doi.org/10.1371/journal.pone.0258214.t005

Table 6. Comparison of independent ROC curves with the model evaluations and the external validations after retraining.

	Difference between areas	SE	95% CI	P-value
Inception3D	0.0185	0.0283	-0.0369 to 0.0740	0.5124
ResNet3D	0.0251	0.0255	-0.0249 to 0.0751	0.3247

https://doi.org/10.1371/journal.pone.0258214.t006

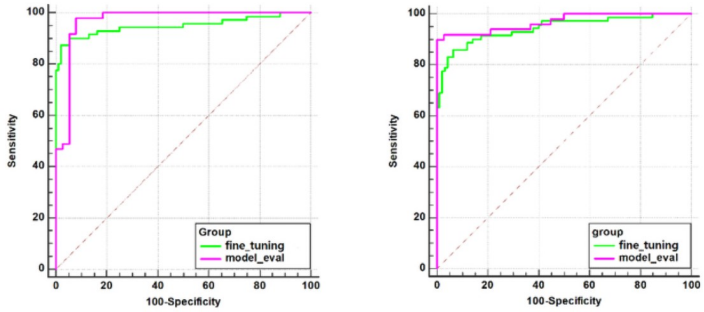

Fig 10. ROC curve comparison between the external validation after retraining and the model evaluation. The '*fine_tuning*' curve indicates the external validation after retraining and '*model_eval*' indicates the model evaluation.

https://doi.org/10.1371/journal.pone.0258214.g010

4 Discussion

Currently, PET scans are used to diagnose AD by determining the level of amyloid deposition, which is assessed for severity through visual assessment by experts. We also assessed it and obtained Fleiss' Kappa coefficient to determine it statistically, and the interreader agreement was derived as $\kappa = 0.8950$. However, this method cannot provide consistent results, as different specialists may interpret images differently. In addition, a doctor's prior experience can have a significant impact on the reliability of the diagnostic results. Therefore, CNN-based medical image analysis, a deep learning procedure, can produce consistent results and improve confidence in the diagnosis.

Similar to other CNN-based neuroimaging classification studies, our model evaluation of the amyloid positivity classification problem in [^{18}F]Florbetaben brain PET yielded an average of approximately 95% accuracy results. In addition, the CNN trained with the single institution dataset demonstrated satisfactory performance when tested with [^{18}F]Florbetaben brain PET images obtained from the subjects in the ADNI database. These results support that this CNN model can help diagnose AD by developing a computer-aided detection tool to determine amyloid positivity since these CNN models can recognize the amyloid deposition features of the brain well.

In the external validation, the classification performance of the VGG 3D model was the best with an AUC of 0.945 and an accuracy of 85.3%, but the optimal model may be different if the experimental conditions are different or other model structures are performed. There was also a difference in the classification performance between the model evaluation and the external validation with the Inception3D and ResNet3D models (significance level < 0.05). Thus, a method to overcome these issues is yet to be developed. This improves optimization by fine tuning [28, 29] weights using the learning part or all the layers in Section 3.4.

4.1 Open platform

Our CNN-based [^{18}F]Florbetaben amyloid brain PET classification studies have a limitation in that only the DAUH dataset and ADNI have been utilized. Therefore, a medical imaging artificial intelligence (AI) research platform that enables doctors with medical knowledge and medical data to perform medical AI research without programming is needed. DEEP:PHI is one such open platform (https://www.deepphi.ai). It is a research platform developed by DEEPNOID, a Korean medical AI startup company that is currently being serviced in the form of a

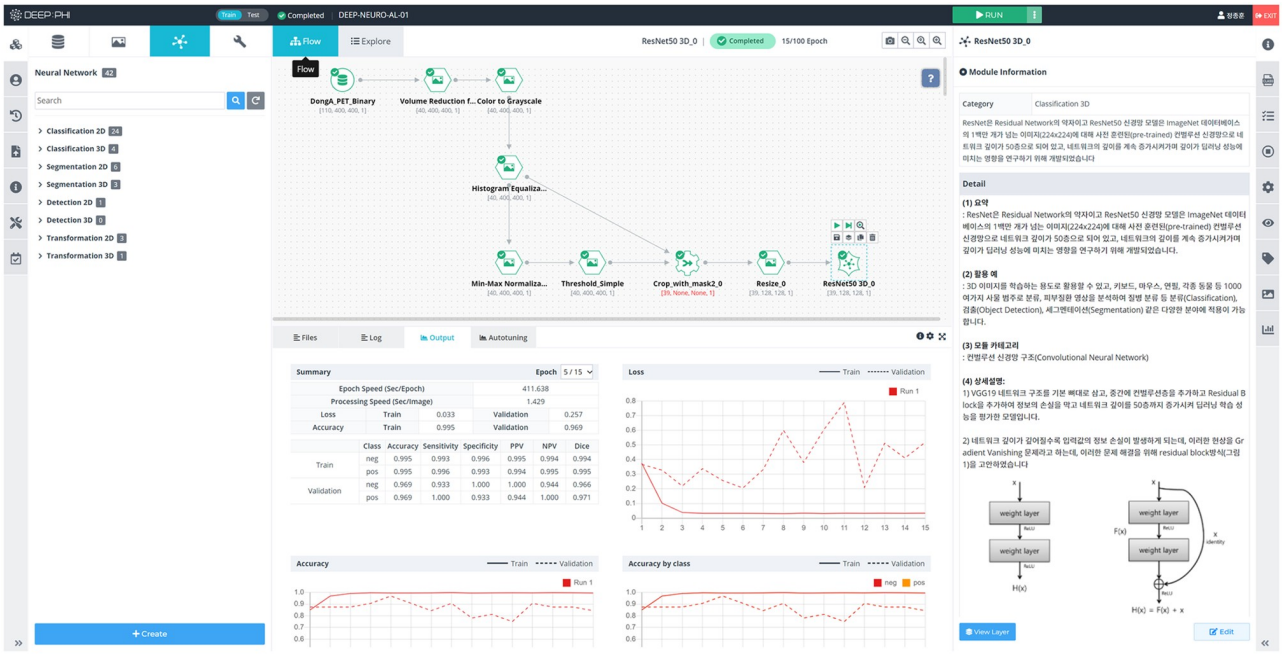

Fig 11. GUI of DEEP:PHI. The image preprocessing, performance, etc. can be seen in the window.

https://doi.org/10.1371/journal.pone.0258214.g011

closed beta, and a number of doctors are using this platform to conduct artificial intelligence research.

As shown in Fig 11, DEEP:PHI comprises a GUI, and it is possible to perform image preprocessing, neural network model generation, and neural network training results verification within the DEEP:PHI platform from the workflow window. In addition, the server provides a high-end research environment without a GPU and hard disk drive in the local environment. As it operates on the web, the platform allows doctors and developers of various organizations to collect data and perform collaborative research directly on the web. The models from our research can also be uploaded to the DEEP:PHI platform and used for various AI research. If necessary, it is possible to modify and create modules specialized for specific research through the code editor.

5 Conclusion

The Aβ classification model was evaluated and external validation was performed with the ADNI dataset. The detailed information on model evaluation and external validation results can be seen in S1 File. The model evaluation results show that the classification performance produces the high accuracy. On the other hand, data from other sources may have differences in quality when compared to the evaluation dataset, which could lead to the poor classification, and preprocessing to minimize these differences is important in external validation. Even though the data have been refined, when the deep learning classification model does not classify tasks well in the external validation step, the model performance can be improved by including other structures or retraining the model. In addition, it is possible to implement an optimal model through various research collaborations using an open platform for medical image AI research.

Supporting information

S1 File. Aβ positive probability using CNN. This is the probability of predicting the label for each CNN model(EXCEL). The label is coded as Aβ positive 1 and Aβ negative 0, and if the probability exceeds 0.5, it is predicted as Aβ positive, and if it is less than 0.5, it is predicted as Aβ negative. One tab is configured for each validation type, and the meaning of the tab name is as follows:

- model_eval: model evaluation.
- external_val: external validation.
- ext_additional: external validation after the fine tunning.
(XLSX)

Acknowledgments

This research was supported by the National Research Foundation (NRF) of Korea funded by the Ministry of Science, ICT & Future Planning (NRF2018 R1A2B2008178).

Author Contributions

Conceptualization: Seung-Yeon Lee, Hyeon Kang, Do-young Kang.

Data curation: Do-young Kang.

Formal analysis: Seung-Yeon Lee, Hyeon Kang.

Funding acquisition: Do-young Kang.

Investigation: Seung-Yeon Lee, Hyeon Kang, Jong-Hun Jeong.

Methodology: Seung-Yeon Lee, Hyeon Kang.

Project administration: Do-young Kang.

Resources: Do-young Kang.

Software: Seung-Yeon Lee, Jong-Hun Jeong.

Supervision: Do-young Kang.

Validation: Seung-Yeon Lee, Hyeon Kang.

Visualization: Seung-Yeon Lee, Hyeon Kang, Jong-Hun Jeong.

Writing – original draft: Seung-Yeon Lee, Jong-Hun Jeong.

Writing – review & editing: Seung-Yeon Lee, Do-young Kang.

References

1. Kang DW, Lim HK. Current Knowledge and Clinical Application of Brain Imaging in Alzheimer's Disease. Journal of Korean Neuropsychiatric Association. 2018; 57(1):12–22. https://doi.org/10.4306/jknpa.2018.57.1.12
2. Varghese T, Sheelakumari R, James JS, Mathuranath PS. A review of neuroimaging biomarkers of Alzheimer's disease. Neurology Asia. 2013; 18(3):239. PMID: 25431627
3. Suk HI, Lee SW, Shen D, Initiative ADN, et al. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage. 2014; 101:569–582. https://doi.org/10.1016/j.neuroimage.2014.06.077 PMID: 25042445

4. Suk HI, Lee SW, Shen D. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Structure and Function. 2015; 220(2):841–859. https://doi.org/10.1007/s00429-013-0687-3 PMID: 24363140
5. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE journal of biomedical and health informatics. 2017; 22(1):173–183. https://doi.org/10.1109/JBHI.2017.2655720 PMID: 28113353
6. Comelli A, Stefano A, Bignardi S, Coronnello C, Russo G, Sabini MG, et al. Tissue classification to support local active delineation of brain tumors. In: Annual Conference on Medical Image Understanding and Analysis. Springer; 2019. p. 3–14.
7. Stefano A, Comelli A, Bravatà V, Barone S, Daskalovski I, Savoca G, et al. A preliminary PET radiomics study of brain metastases using a fully automatic segmentation method. BMC bioinformatics. 2020; 21 (8):1–14. https://doi.org/10.1186/s12859-020-03647-7 PMID: 32938360
8. Cho K, Kim WG, Kang H, Yang GS, Kim HW, Jeong JE, et al. Classification of 18 F-Florbetaben Amyloid Brain PET Image using PCA-SVM. Biomedical Science Letters. 2019; 25(1):99–106. https://doi.org/10.15616/BSL.2019.25.1.99
9. Kang H, Kim WG, Yang GS, Kim HW, Jeong JE, Yoon HJ, et al. VGG-based BAPL score classification of 18F-florbetaben amyloid brain PET. Biomedical Science Letters. 2018; 24(4):418–425. https://doi.org/10.15616/BSL.2018.24.4.418
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.
11. Suzuki K. Overview of deep learning in medical imaging. Radiological physics and technology. 2017; 10 (3):257–273. https://doi.org/10.1007/s12194-017-0406-5 PMID: 28689314
12. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical image analysis. 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
13. Liu KL, Wu T, Chen PT, Tsai YM, Roth H, Wu MS, et al. Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation. The Lancet Digital Health. 2020; 2(6):e303–e313. https://doi.org/10.1016/S2589-7500(20)30078-9 PMID: 33328124
14. Baldwin DR, Gustafson J, Pickup L, Arteta C, Novotny P, Declerck J, et al. External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. Thorax. 2020; 75(4):306–312. https://doi.org/10.1136/thoraxjnl-2019-214104 PMID: 32139611
15. Ashburner J, Friston KJ. Nonlinear spatial normalization using basis functions. Human brain mapping. 1999; 7(4):254–266. https://doi.org/10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G PMID: 10408769
16. Schmidt ME, Matthews D, Andrews R, Mosconi L. Positron emission tomography in Alzheimer disease: diagnosis and use as biomarker endpoints. In: Translational Neuroimaging. Elsevier; 2013. p. 131–174.
17. Win TP, Hosokai Y, Minagawa T, Muroi K, Miwa K, Maruyama A, et al. Comparison of Count Normalization Methods for Statistical Parametric Mapping Analysis Using a Digital Brain Phantom Obtained from Fluorodeoxyglucose-positron Emission Tomography. Asia Oceania Journal of Nuclear Medicine and Biology. 2019; 7(1):58. https://doi.org/10.22038/AOJNMB.2018.11745 PMID: 30705912
18. Kalavathi P, Prasath VS. Methods on skull stripping of MRI head scan images—a review. Journal of digital imaging. 2016; 29(3):365–379. https://doi.org/10.1007/s10278-015-9847-8 PMID: 26628083
19. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 1–9.
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818–2826.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
22. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR; 2015. p. 448–456.
23. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. In: ICML; 2010. p. 807–814. Available from: https://icml.cc/Conferences/2010/papers/432.pdf.
24. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014; 15(1):1929–1958.
25. Hosseini-Asl E, Keynton R, El-Baz A. Alzheimer's disease diagnostics by adaptation of 3D convolutional network. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE; 2016. p. 126–130.

26. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:181112808. 2018;.

27. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008; 9(11).

28. Hussain M, Bird JJ, Faria DR. A study on cnn transfer learning for image classification. In: UK Workshop on computational Intelligence. Springer; 2018. p. 191–202.

29. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009; 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191