Biomedical Science Letters 2018, 24(4): 418~425
https://doi.org/10.15616/BSL.2018.24.4.418
eISSN : 2288-7415

Original Article

VGG-based BAPL Score Classification of 18F-Florbetaben Amyloid Brain PET

Hyeon Kang[1,*], Woong-Gon Kim[3,], Gyung-Seung Yang[5,**], Hyun-Woo Kim[4,**], Ji-Eun Jeong[1,2,**], Hyun-Jin Yoon[1,2,**], Kook Cho[1,**], Young-Jin Jeong[1,2,***] and Do-Young Kang[1,2,†,***]**

[1]Institute of Convergence Bio-Health, Dong-A University, Busan 49201, Korea
[2]Department of Nuclear Medicine, Dong-A University Medical Center, Dong-A University College of Medicine, Busan 49201, Korea
[3]Economic Survey, Gyeongin Regional Statistics Office, Gwacheon 13809, Korea
[4]Department of Industrial Engineering, Hanyang University, Seoul 04763, Korea
[5]Ubicod Company, Seoul 08381, Korea

Amyloid brain positron emission tomography (PET) images are visually and subjectively analyzed by the physician with a lot of time and effort to determine the β-Amyloid (Aβ) deposition. We designed a convolutional neural network (CNN) model that predicts the Aβ-positive and Aβ-negative status. We performed 18F-florbetaben (FBB) brain PET on controls and patients (n=176) with mild cognitive impairment and Alzheimer's Disease (AD). We classified brain PET images visually as per the on the brain amyloid plaque load score. We designed the visual geometry group (VGG16) model for the visual assessment of slice-based samples. To evaluate only the gray matter and not the white matter, gray matter masking (GMM) was applied to the slice-based standard samples. All the performance metrics were higher with GMM than without GMM (accuracy 92.39 vs. 89.60, sensitivity 87.93 vs. 85.76, and specificity 98.94 vs. 95.32). For the patient-based standard, all the performance metrics were almost the same (accuracy 89.78 vs. 89.21), lower (sensitivity 93.97 vs. 99.14), and higher (specificity 81.67 vs. 70.00). The area under curve with the VGG16 model that observed the gray matter region only was slightly higher than the model that observed the whole brain for both slice-based and patient-based decision processes. Amyloid brain PET images can be appropriately analyzed using the CNN model for predicting the Aβ-positive and Aβ-negative status.

Key Words: Alzheimer's disease, β-Amyloid, Convolutional neural network, 18F-florbetaben PET, Gray matter

INTRODUCTION

β-Amyloid (Aβ) is considered an important hallmark required to understand Alzheimer's disease (AD) and predict disease prognosis (Hardy et al., 1991; Gunasekaran et al., 2015). 18F-florbetaben (FBB) is a type of Aβ-targeting radiopharmaceutical tracer. 18F-FBB amyloid positron emission tomography (PET) provides accurate and early diagnosis with high sensitivity and specificity (Barthel et al., 2011).

Received: November 26, 2018 / Accepted: December 6, 2018
[*]Graduate student, [**]Researcher, [***]Professor.
[†]Corresponding author: Do-Young Kang. Department of Nuclear Medicine, Dong-A University Medical Center, Dong-A University College of Medicine, #26 Daesingongwon-ro, Seo-gu, Busan, 49201, Korea.
Tel: +82-51-240-5630, Fax: +82-51-242-7237, e-mail: dykang@dau.ac.kr

Table 1. Demographic details of the patients used to train/validate the selected model

Characteristics	BAPL*1	BAPL 2	BAPL 3	Total
Patients (n*)	60	53	60	173
Mean of age	67.55	73.02	69.08	69.76
SD* of Age	9.01	5.64	8.83	8.33
Female (F/Total)	38 (0.63)	32 (0.60)	27 (0.45)	97 (0. 56)
No. slice of BAPL 1	2160	931	0	3,091
No. slice of BAPL 2	0	977	0	977
No. slice of BAPL 3	0	0	2,160	2,160

*BAPL: brain amyloid plaque load; n: number; SD: standard deviation

Brain amyloid plaque load (BAPL) score a is pre-defined three-grade scoring system wherein measurements are made by the physician according to the visual assessment of the subject's amyloid deposition in the brain using 18F-FBB. BAPL scores of 1 (BAPL 1), 2 (BAPL 2), and 3 (BAPL 3) indicate no Aβ load, minor Aβ load, and significant Aβ load, respectively. Therefore, BAPL 1 is considered to indicate Aβ-negative status, whereas BAPL 2 and BAPL 3 indicate Aβ-positive status (Barthel et al., 2011).

The visual evaluation of the image by the clinician is the most reliable way of image evaluation; however, it involves the disadvantage of being time-consuming and labor-intensive. Moreover, a numerical comparison is impossible, resulting in inter-observer problems (Gulshan et al., 2016). In conventional amyloid PET image analysis, comparison and staging through pre-defined cutoffs using metrics, such as standard uptake value ratio, and statistical analysis techniques using statistical parametric mapping (SPM) were treated as quantitative methods to resolve these issues (Lopresti et al., 2005). Recently, new image analysis technologies that use deep learning have been applied to the field of medicine in medical imaging, and remarkable achievements have been reported (Gulshan et al., 2016; Lakhani et al., 2017).

In the present study, we compared the classification performance of the slice-based posterior probabilities of Aβ-positive status, calculated using the pre-trained VGG16 model as per gray matter masking obtained with only functional PET and the estimated subject-based posterior probabilities using rule-based approach to mimic the clinical setting.

MATERIALS AND METHODS

Subjects

All the data used in this study were retrospectively sampled at the Department of Nuclear Medicine, Dong-A university hospital (DANM), from November 2015 to May 2018. The study population involved totally 173 subjects; 60 subjects had a BAPL score of 1, 53 had a score of 2, and 60 had a score of 3. Detailed information of the study population is presented in Table 1.

Amyloid PET dataset

Image labeling and sampling: Clinical information, including the qualitative analysis of 18F-FBB PET and BAPL scores of the DANM dataset, was arranged in cooperation with Department of Neurology, Dong-A university hospital.

According to the current diagnostic criteria for 18F-FBB PET, BAPL 1 is considered a state of Aβ-negative, whereas BAPL 2 and BAPL 3 are considered to indicate Aβ-positive status. Such a decision depends on the visual assessment by the clinician in the trans-axial plane. Therefore, we indexed the slices in the trans-axial plane for each patient data; therefore, we performed a visual assessment after sampling the pre-determined slices (total 36 slices from 15^{th} to 50^{th} slices from 68 slices per a person). Fig. 1 shows an indexed anatomical slice of randomly sampled patient data.

In this study, we considered BAPL 2 and BAPL 3 as constituting one class according to the definition of BAPL consensus to implement the convolutional neural network (CNN) model for discriminating the Aβ-positive and Aβ-

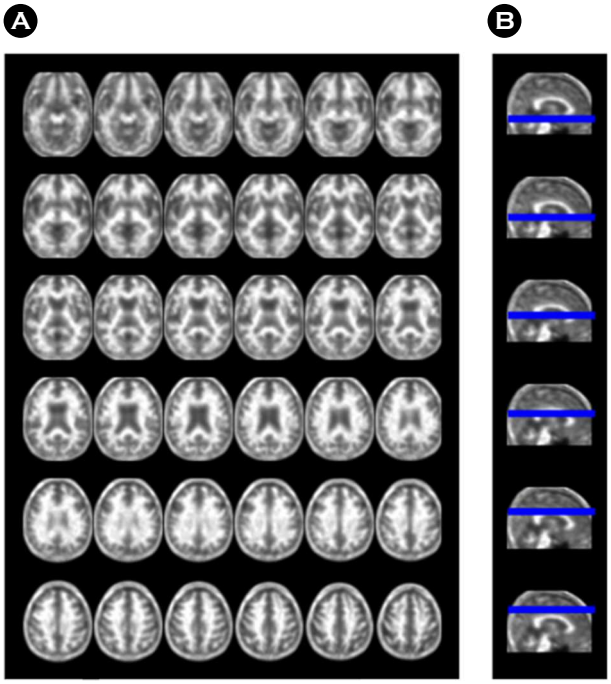

Fig. 1. The 36 images sampled using 18F-FBB amyloid PET on the trans-axial plane used to train the VGG16 model. (A) shows 36 slices sampled from the 15th to 50th slices from 68 slices on the trans-axial plane. Each slice was indexed in the dorsal direction from the bottom to the top in a brain and arranged on 6 × 6. (B) shows the level of the brain placed in each row of the (A).

negative status; therefore, the final problem is defined as a binary classification.

Data Pre-processing: The voxel size of raw PET images after acquisition from the scanner was 400 × 400 × 110 (height × width × depth). The raw PET images were spatially normalized using the normalization module from SPM8. The whole sample dataset was non-linearly registered to the PET template; that was created using 21 NC and 9 AD subjects who imaged 18F-FBB Amyloid PET. The resulting images obtained from SPM8 have a 3-dimensional voxels image size of 95 × 79 × 68.

We performed an experiment that only considers the regional information on gray matter as much as possible by masking each sample using the gray matter masking (GMM) obtained from the PET template used to register the entire data and the volume of the whole brain. Fig. 3 shows the state and difference in the images observed in this experimental procedure.

Following this, all the data were normalized such that the mean was 0 and the standard deviation was 1 prior to entering a selected model. All the above procedures were performed outside the outer and inner loop of the nested cross validation (NCV). In the NCV, the inner loop operates a wrapper algorithm that searches for hyper-parameters by performing hold-out again on the subset from the entire dataset for hyper-parameter validation, such as grid search or random search algorithm (Pedregosa et al., 2011; Bergstra et al., 2012; Taylor et al., 2017). In the present study, we used Bayesian optimization (Snoek et al., 2012) as the wrapper algorithm. In each outer loop, the hyper-parameter ultimately determined from each inner loop is used to select the best model for estimating the generalization performance in the outer loop (Varma et al., 2006).

In order to select the appropriate deep learning model and test it with the above data set, we conducted 4-fold NCV and performed stratified sampling to generate each fold. We determined the number of folds as observable performance variables while maintaining appropriate variance in the limited data set.

Finally, data augmentation was performed on each BAPL group up to 3000 slices for each number of data on each group to become equal by flipping and rotation method to make a selected model learn a positional invariance and give BAPL 2 group, which was relatively considered as a small data set, more opportunities to be sampled as a batch sample to be used in the optimization of the model parameters by increasing the number of data belonging to BAPL 2 group. Furthermore, data augmentation is a process of learning more about robust features in the model by learning about the positional invariance of the object in the sample. Therefore, all the augmentation processes used in our experiments were performed on each loop of NCV. All the data pre-processing after spatial normalization is written using packages implemented using the Python source code. (Python 3.5.2, imgaug 0.2.6)

Convolutional neural network

The VGG16 (visual geometry group) model demonstrated the effect of the deep network structure created using the 3

Table 2. Comparison of the classification performance according to pre-process (%) (SD)

	Accuracy	Aβ-negative recall	Aβ-positive recall
Slice based classification without GMM*	89.60 (1.77)	95.32 (2.75)	85.76 (3.29)
Slice based classification with GMM	92.39 (2.51)	98.94 (0.95)	87.93 (4.36)
Subject based classification without GMM	89.21 (1.14)	70.00 (3.85)	99.14 (1.72)
Subject based classification with GMM	89.78 (6.01)	81.67 (17.53)	93.97 (5.89)

*GMM: gray matter masking

Table 3. Hyper-parameter space searched using Bayesian optimization used in all the inner loops

Type	without GMM*	with GMM	Total
Learning rate	0.0007 (0.0017)	0.0013 (0.0026)	0.0011 (0.0023)
No. epoch	15.25 (25.91)	23.25 (27.51)	20.58 (26.69)
No. hidden nodes	1 (0.93)	1 (0.52)	1 (0.66)
No. hidden layers	158 (117.41)	185.38 (149.51)	176.25 (127.65)
Accuracy (%) (SD)	99.69 (0.88)	97.50 (6.58)	98.2 (5.44)

GMM: gray matter masking

× 3 size of filters in ImageNet competition (Simonyan et al., 2014) and is also adopted as a baseline in experimental comparison with many modified models (Long et al., 2015; He et al., 2016). In order to classify the pre-processed amyloid PET voxel data, we selected the VGG16 model as a fixed classifier for our experiment and attempted to estimate the posterior probability for the defined binary label. Here the parameter of the VGG16 model was used to derive better generalization performances via transfer learning by referring to the model parameters learned in the data set of the ImageNet competition from the Keras library (Keras 2.2.2 version). The hyper-parameters were searched through a Bayesian optimization and validation was performed in the inner loop of the NCV. In the outer loop of the NCV, the generalization performance for the model wherein the hyper-parameter was determined in the inner loop was estimated. The parameter space searched with Bayesian optimization was the learning rate from 1e-7 to 1e-2, the number of epochs from 5 to 100, the number of hidden nodes on a dense layer from 5 to 512 and the number of depths for hidden layers of fully connected layer from 0 to 3. Prior to the statistical estimation via Bayesian optimization, the initial candidate was 1e-5 (learning rate), 5 (No. epoch), 1 (No. hidden node), and 128 (No. hidden node) and arbitrarily determined by the practitioner. Specific details were given in Table 2. The model parameters of the VGG16 model determined in each NCV loop were converged through the Adam optimizer and back propagation algorithm implemented in the Keras library and their default values.

In the present study, we ultimately estimated the output as a posterior probability for a sample to be Aβ-positive from the slice-based standard to the patient-based standard using our selected VGG model. The slice-based Aβ estimation was calculated using the VGG16 model and the patient-based estimation was determined to be positive for at least one positive output from a slice-based decision on the probability obtained through the VGG16 model by mimicking the situation in actual clinical practice.

RESULTS

Table 3 shows the average and standard deviation values of each performance estimated from the outer loop of 4-fold NCV as per the observed brain region and evaluation standards. In order to compare the performance as per GMM at a slice-based standard, the performance of the model that only observed the gray matter region was higher on all the performance metrics as compared to that of the model ob-

serving the whole brain region (accuracy 92.39 vs. 89.60, sensitivity 87.93 vs. 85.76, and specificity 98.94 vs. 95.32). In contrast, the performances as per masking at a patient-based standard were almost similar (89.78 vs. 89.21), especially in terms of the accuracy; however, the sensitivity when the GM region was included was lower compared to when the white matter region was included (93.97 vs. 99.14). Further, the specificity on inclusion of the GM region was higher than that when the white matter region was included (81.67 vs. 70.00).

In order to evaluate the discrimination power of the selected model that estimates the slice-level posterior probability and the rule-based estimation approach, the area under curve (AUC) was calculated using receiver operating characteristic (ROC) analysis for each experiment. The AUC when the VGG16 model observed only the gray matter region was slightly higher as compared to when the whole brain was observed for both the slice-based and patient-based decision

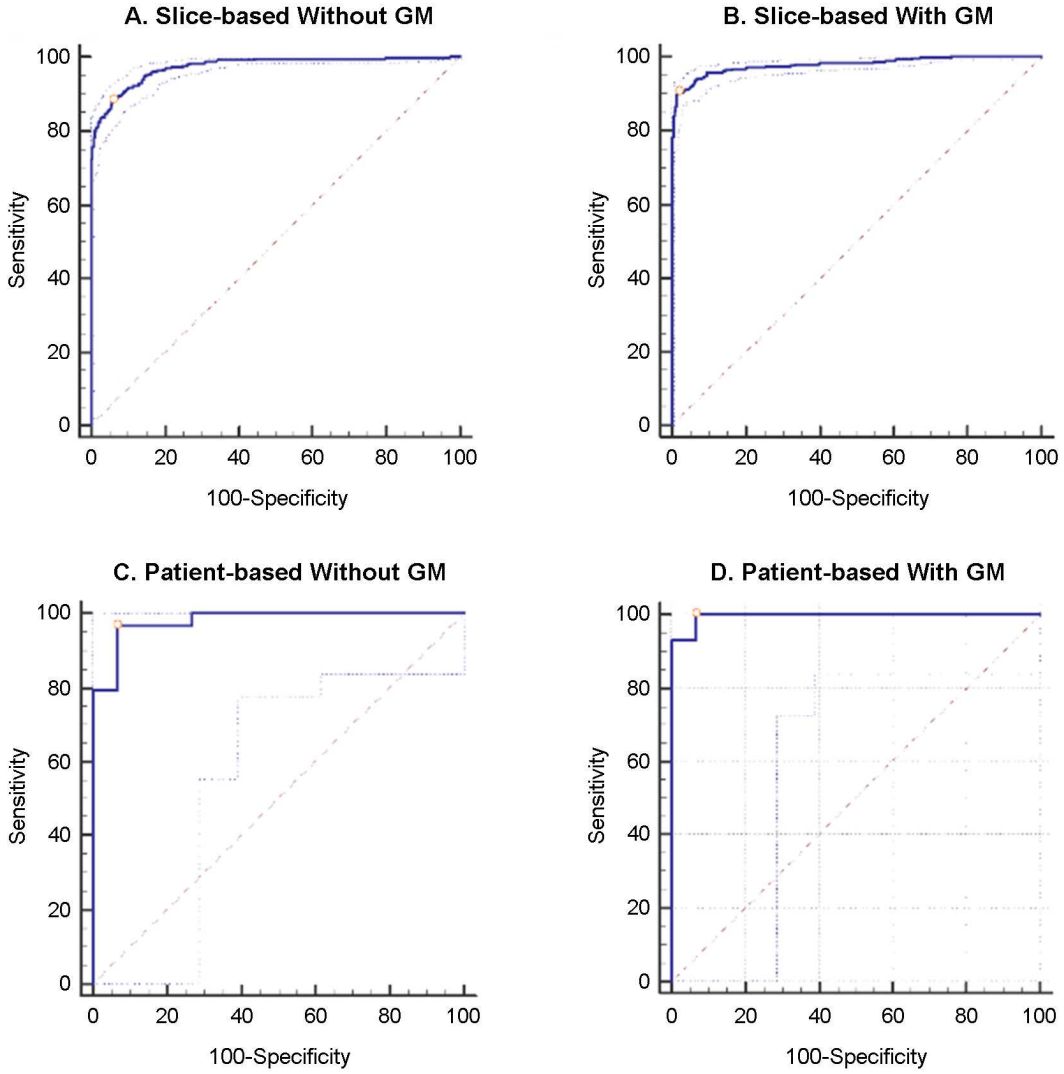

Fig. 2. Receiver operating characteristic (ROC) analysis for the classification on several experiments slice-based/patient-based and with/without gray matter masking. The dotted line on the figure shows 95% confidence interval of the ROC curves. A and B show the results of the classification performance that the selected model estimates in a slice-based standard according to the masking gray matter. C and D show the ROC curve of the rule-based classification that mimics the condition in actual clinical practice using slice-based estimation from the A and B processes. Values for the area under the curve for A, B, C, and D were 0.972, 0.976, 0.973, and 0.983, respectively.

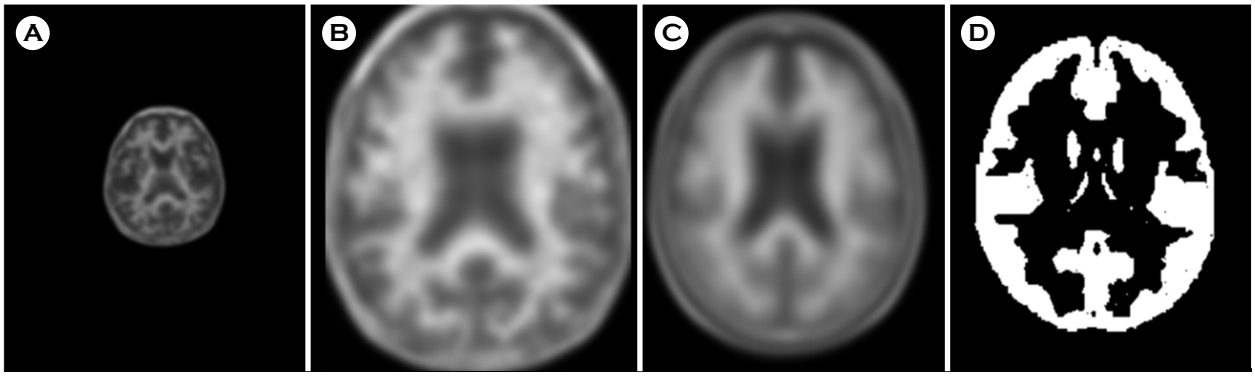

Fig. 3. Amyloid PET images used in this study. Fig. 3-A shows the raw PET slice after acquisition. Fig. 3-B is the cropped image and registered into Fig. 3-C space of the 18F FBB PET template. And Fig. 3-D is mask of gray matter obtained from Fig. 3-C (threshold = 0.5).

process (Fig. 2). A and B show the results of the classification performance that the selected model estimated in the slice-based standard according to masking gray matter. C and D show the ROC of the rule-based classification, mimicking the situation in an actual clinical setting using slice-based estimations from A and B processes. AUC for A, B, C, and D were 0.972, 0.976, 0.973, and 0.983, respectively.

Hyper-parameters determined from an inner loop of the NCV and used for the estimation of slice-based classification performance are summarized in Table 3. In the search candidates of the Bayesian optimizer, hyper-parameter values were higher than those while considering the whole brain, except the hidden nodes. Moreover, comparing the performance estimation in each inner loop in the NCV applied to the experiment, the variance of the estimated performance is less than the variance of the hyper-parameters.

DISCUSSION

In the conventional quantitative analysis of amyloid PET as per the pathological deposition patterns of amyloid plaque, studies have considered a region of gray matter as the region of interest (Choi et al., 2016). However, the deep learning approach is a process of learning a mathematical function with millions of parameters in a data set and finding a meaningful feature from the input data. This approach has been applied for various purposes in the field of medicine with remarkable achievements (Gulshan et al., 2016). Our experimental results show that both, the sensitivity and specificity are improved when a pre-trained VGG16 model is learned by inputting the data of the gray matter separately from that of the whole brain in the slice-based classification. This may indicate that the features obtained by observing only the gray matter involve more distant and discriminating features between the Aβ-positive and Aβ-negative groups than those obtained by observing the whole brain.

In addition, in our results, the pre-trained VGG16 showed about 90% accuracy, suggesting that the method of generating amyloid PET template for brain spatial normalization and mask for gray matter using only the functional image may be enough for performing quantitative analysis. This suggests that it can be applicable as a quantitative method in cases where the anatomical image is not available. Moreover, even though the count normalization for each lobe of the reference region, such as cerebellar gray matter, was not applied in our data set, the pre-trained VGG16 model showed a discriminating performance, including visual evaluation of a human clinician using the contrast of activity rather than the count of activity.

The rule of subject-based determinations from the slice-based output of the pre-trained VGG16 model was considering the highest posterior probability from a list of 36 slice-based posterior probabilities for each subject to be Aβ-positive as an ultimate posterior probability for a subject. Therefore, the false-positive rate for each slice exerts a considerable effect on the false-positive rate of the subject-based

decision. Although the sensitivity of the subject-based decision without GMM was higher than that of the subject-based decision with GMM, the specificity was lower. In the entire outer loop of the NCV, 18 out of 19 cases were incorrectly identified with BAPL 1, and 1 case was incorrectly identified with BAPL 2 for the subject-based experiment without gray matter; 11 out of the 18 cases were incorrect BAPL 1, and 7 other cases were identified with BAPL 2 for the subject-based experiment with gray matter. Therefore, further research is required in a follow-up study on how to distinguish between BAPL 1 and BAPL 2 rather than BAPL 3.

We trained the pre-trained VGG16 model to estimate the posterior probabilities of Aβ-positive for each slice of a subject. Moreover, the posterior probabilities for each subject were estimated using the calculated posterior probabilities, and we evaluated the performance as a predictive model. In particular, we performed pre-processing, including spatial normalization and acquisition of GMM using only functional images. Finally, we compared the performance of the classifiers considering only the gray matter or the whole brain. The data set used in the experiment was that of 173 subjects imaged using 18F-FBB PET. We found that the slice-based classification with GMM showed better performance for distinguishing between the Aβ-positive and Aβ-negative groups.

ACKNOWLEDGEMENT

This research was supported by the project at Institute of Convergence Bio-Health, Dong-A University funded by Busan Institute of S&T Evaluation and Planning.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Barthel H, Gertz HJ, Dresel S, Peters O, Bartenstein P, Buerger K, Hiemeyer F, Wittemer-Rump SM, Seibyl J, Reininger C, Sabri O. Cerebral amyloid-β PET with florbetaben (18F) in patients with Alzheimer's disease and healthy controls: a multicenter phase 2 diagnostic study. The Lancet Neurology. 2011. 10: 424-435.

Bergstra J, Bengio Y. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 2012. 13: 281-305.

Choi WH, Um YH, Jung WS, Kim SH. Automated quantification of amyloid positron emission tomography: a comparison of pmod and mimneuro. Annals of Nuclear Medicine. 2016. 30: 682-689.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams Tom, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016. 316: 2402-2410.

Gunasekaran TI, Ohn T. MicroRNAs as Novel Biomarkers for the Diagnosis of Alzheimer's Disease and Modern Advancements in the Treatment. 2015. Biomedical Science Letters. 2015. 21: 1-8.

Hardy J, Allsop D. Amyloid deposition as the central event in the aetiology of Alzheimer's disease. Trends in Pharmacological Sciences. 1991. 12: 383-388.

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. 770-778.

Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017. 284: 574-582.

Long J, Shelhamer E, Darrel T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. 3431-3440.

Lopresti BJ, Klunk WE, Mathis CA, Hoge JA, Ziolko SK, Lu X, Meltzer CC, Schimmel K, Tsopelas ND, DeKosky ST, Price JC. Simplified quantification of Pittzburgh Compound B amyloid imaging PET studies: a comparative analysis. Journal of Nuclear Medicine. 2005. 46: 1959-1972.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in python. Journal of Machine Learning Research. 2011. 12: 2825-2830.

Simonyan K, Zisserman A. Very Deep convolutional networks for large-scale image recognition. arXiv preprint. 2014. arXiv: 1409-1556.

Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization

of machine learning algorithm. In Advances in Neural Information Processing System. 2012. 2951-2959.

Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123) FP-CIT classification: the beginning of the end for semi-quantification?. EJNMMI Physics. 2017. 4: 29.

Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006. 7: 91.

https://doi.org/10.15616/BSL.2018.24.4.418

Cite this article as: Kang H, Kim WG, Yang GS, Kim HW, Jeong JE, Yoon HJ, Cho K, Jeong YJ, Kang DY. VGG-based BAPL Score Classification of 18F-Florbetaben Amyloid Brain PET. Biomedical Science Letters. 2018. 24: 418-425.