



Comparison of Three Automated Approaches for Classification of Amyloid-PET Images

Ying-Hwey Nai¹ · Yee-Hsin Tay² · Tomotaka Tanaka^{3,4} · Christopher P. Chen^{4,5} · Edward G. Robins^{1,6} · Anthonin Reilhac¹ · for the Alzheimer's Disease Neuroimaging Initiative

Accepted: 25 April 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Automated amyloid-PET image classification can support clinical assessment and increase diagnostic confidence. Three automated approaches using global cut-points derived from Receiver Operating Characteristic (ROC) analysis, machine learning (ML) algorithms with regional SUV_r values, and deep learning (DL) network with 3D image input were compared under various conditions: number of training data, radiotracers, and cohorts. 276 [¹¹C]PiB and 209 [¹⁸F]AV45 PET images from ADNI database and our local cohort were used. Global mean and maximum SUV_r cut-points were derived using ROC analysis. 68 ML models were built using regional SUV_r values and one DL network was trained with classifications of two visual assessments – manufacturer's recommendations (gray-scale) and with visually guided reference region scaling (rainbow-scale). ML-based classification achieved similarly high accuracy as ROC classification, but had better convergence between training and unseen data, with a smaller number of training data. Naïve Bayes performed the best overall among the 68 ML algorithms. Classification with maximum SUV_r cut-points yielded higher accuracy than with mean SUV_r cut-points, particularly for cohorts showing more focal uptake. DL networks can support the classification of definite cases accurately but performed poorly for equivocal cases. Rainbow-scale standardized image intensity scaling and improved inter-rater agreement. Gray-scale detects focal accumulation better, thus classifying more amyloid-positive scans. All three approaches generally achieved higher accuracy when trained with rainbow-scale classification. ML yielded similarly high accuracy as ROC, but with better convergence between training and unseen data, and further work may lead to even more accurate ML methods.

Keywords Alzheimer's disease · Positron emission tomography (PET) · Visual interpretation · Equivocal · Machine Learning · Deep Learning

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease defined by the abnormal deposits of amyloid-beta (A β) plaques and neurofibrillary tau tangles in the brain. Positron emission tomography (PET) imaging with A β -targeting radiotracers is a crucial tool for the in vivo observation and the quantitative measurements of A β burden (Jack et al., 2018; Johnson et al., 2013). These measurements are used in research to investigate the fundamental biological mechanisms involving A β and their interaction with concomitant conditions and to assess the disease severity, progression, and the effect of therapeutics over time. A β -positivity status based on the visual assessment of A β -PET scans plays a central role in the confirmation of clinical diagnosis, and also for subject selection in research studies or clinical trials of A β -targeting

Ying-Hwey Nai and Anthonin Reilhac made an equal contribution to this manuscript.

Alzheimer's Disease Neuroimaging Initiative (ADNI) is a Group/Institutional Author.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

✉ Ying-Hwey Nai
mednyh@nus.edu.sg; yinghweynai@yahoo.com

Extended author information available on the last page of the article

therapeutics. Following visual assessment, a single cut-point value, separating abnormal scans from normal scans based on their global SUVR values, is usually determined using Receiver Operating Characteristics (ROC). These cut-points can be then used to automatically classify subsequent scans to support/replace the subjective, tedious and resource-intensive visual assessments.

Artificial intelligence (AI) based methods have recently gained attention as alternative ways for scan classification. With these approaches, the decision is either made from radiotracer uptake values measured in typical AD cortical regions with Machine Learning (ML) models (Kim et al., 2020) or directly from the 3D image with Deep Learning (DL) models (Kang et al., 2018; Son et al., 2020). However, little effort has been done to assess the predictive values of these classification methods and to determine the required number of training scans to build models that are generalizable to subsequent scans. In addition, the quality of the fully qualified scans that are used to build/train models is of paramount importance but is often overlooked. In the case of A β -PET scans, despite the use of established procedures to ensure consistency and accuracy (Rowe & Villemagne, 2013), classification outcomes and inter-rater agreements may differ with color-scales, visualization program, assessment criteria, image intensity scaling method, and differences in image quality (Lundeen et al., 2018) possibly leading to inconsistency in the training data with unknown impact on the model performance. Finally, A β -targeting radiotracers differ by their binding affinities to A β plaques in the gray matter (GM) and by their level of nonspecific binding, mainly to myelin in the white matter (WM), adding another level of complexity and source of discrepancy in rater's assessments considering that visual rating mainly relies on image contrast.

We conducted this work with the hypothesis that AI technology should offer more accurate surrogate models of visual interpretation leading to higher classification accuracy than ROC classification for both [¹¹C]PiB and fluorinated amyloid-PET radiotracers, specifically [¹⁸F]AV45. We investigated and compared the predictive values of cut-points derived from ROC analysis, as well as of ML and DL methods under various conditions: radiotracers, and cohorts. Furthermore, we built the training datasets using two well-established visual assessment methods and both

homogeneous and heterogeneous datasets, from local cohort acquired at our center and multicenter data from publicly-available database, in order to investigate their impact on the performance of the automated classification methods. Post mortem tissue samples would provide the ideal ground truth for A β -positivity classification. However, our work is concerned with replicating human assessment of PET images for diagnosis while the patient is still alive.

Material and Methods

PET Image Data

A total of 103 [¹¹C]PiB and 209 [¹⁸F]AV45 ([¹⁸F]Florbetapir / [¹⁸F]Amyvid) processed baseline PET scans were obtained from the ADNI database (www.adni-info.org). The clinical and demographics characteristics of subjects can be found in Table 1. The processing steps included motion correction, intensity normalization using a subject-specific mask so that the average of voxels within the mask is exactly one, spatial normalization, and resampling into a common space of 160 × 160 × 96 matrix with 1.5 mm cubic voxels. PET scans originated from different scanners and images were also subsequently filtered with a scanner-specific filter function to produce images with a uniform isotropic resolution of 8 mm (full width half maximum). In addition, 176 [¹¹C]PiB PET images were selected from a local cohort, which was recruited with the primary aim of gaining novel insights into the joint effects of brain A β burden and cerebral small vessel diseases on neurodegeneration and cognition, including thus a significant proportion of subjects with concomitant cerebrovascular diseases (CeVD). These [¹¹C]PiB PET images were acquired on the Biograph mMR (Siemens Healthcare GmbH) at our institute, the Clinical Imaging Research Centre (CIRC), in conjunction with the Memory Aging and Cognition Centre (MACC) at the National University of Singapore. Written informed consent was obtained in the preferred language of the participants or accompanying relatives. Ethics approval was obtained from the National-Healthcare Group Domain-Specific Review Board and the study was conducted following the Declaration of Helsinki.

List-mode data were acquired for 30 min, 40 min after the intravenous injection of 370 (\pm 15%) MBq of [¹¹C]PiB

Table 1 Clinical and demographic characteristics of subjects

	ADNI-[¹⁸ F]AV45	ADNI-[¹¹ C]PiB	CIRC-[¹¹ C]PiB
Age	73 \pm 7.6 (56–94)	75.7 \pm 7.7 (55–90)	75.6 \pm 7.24 (54–92)
Gender (M/F)	121 / 88	67 / 36	80/96
Clinical Diagnosis	37 CN, 151 MCI, 21 AD	19 CN, 65 MCI, 19 AD	29 CN, 97 MCI, 17 VAD, 33 AD
MMSE scores	27.7 \pm 2.4 (19–30)	26.5 \pm 3.0 (15–30)	22.7 \pm 5.5 (6–30)

Values represent the mean \pm stdev (min – max)

CN Cognitive normal, MCI mild cognitive impaired, AD Alzheimer's disease, VAD Vascular Dementia

(Tanaka et al., 2020). Motion correction was applied using an in-house developed rebinner (Reilhac et al., 2018) during the framing of the list-mode into a single static frame that was then reconstructed with all corrections, including resolution modeling, into a $344 \times 344 \times 127$ image matrix with a voxel size of $2.09 \text{ mm} \times 2.09 \text{ mm} \times 2.03 \text{ mm}$ using 3 iterations/21 subsets of 3D ordinary Poisson ordered-subsets expectation–maximization (Tanaka et al., 2020). Images were then registered to the same space as the ADNI datasets, to obtain $160 \times 160 \times 96$ matrices with 1.5 mm cubic voxels.

Visual Assessments

Two visual assessment methods—rainbow-scale and gray-scale, named after the color-scales employed, were used to classify each scan (Fig. 1). The gray-scale followed the manufacturer’s recommended methods for $[^{18}\text{F}]\text{AV45}$, whereby images are assessed using the inverted gray color-scale with the image intensity adjusted to obtain the highest contrast between GM and WM (Eli Lilly, 2019). The rainbow-scale is a combination of commonly-applied methods for $[^{11}\text{C}]\text{PiB}$, using the rainbow-scale with the image intensity adjusted until the cerebellar WM is largely yellow with a few small spots turning red (Ng et al., 2007; Tanaka et al., 2020; Yamane et al., 2017). $\text{A}\beta$ -positivity required a minimum of one AD specific cortical region to be positive which translated into reddish color with the rainbow-scale or loss of GM-WM demarcation with the gray-scale. The cortical regions included the frontal lobe, parietal lobe, temporal lobe, anterior cingulate, posterior cingulate, and precuneus.

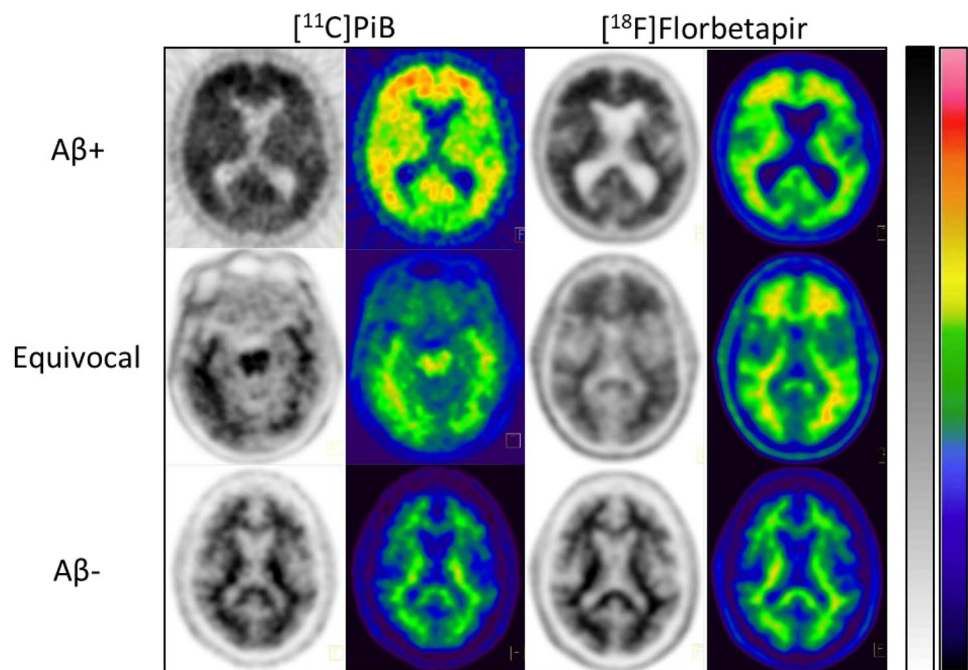
Scans were classified as $\text{A}\beta+$, $\text{A}\beta-$ or equivocal. However, for the final binary classification, equivocal cases were assigned either as $\text{A}\beta+$ or $\text{A}\beta-$ upon the readers’ consensus decision. All images were visualized in the 16-bit color-scale on the Syngo platform (Siemens AG, Germany) in the transaxial orientation, but also in the sagittal and coronal planes when needed, and were viewed systematically from the bottom to the top of the brain.

The $\text{A}\beta$ -positivity was independently assessed by an inexperienced student trained in the visual assessment of $\text{A}\beta$ -PET scans, a junior PET image analysis researcher, a senior neuro-PET researcher, and an experienced medical doctor. Readers evaluated all the $[^{18}\text{F}]\text{AV45}$ and $[^{11}\text{C}]\text{PiB}$ images using the rainbow-scale before the gray-scale, with more than 2 weeks apart to prevent readers’ memory from affecting the assessments. The final consensus was achieved subsequently by the medical doctor and senior neuro-PET researcher. Differences in classification between the two procedures were evaluated with Cohen’s Kappa coefficient (κ) (McHugh, 2012) for each radiotracer. Inter-rater agreement among the 3 raters (junior PET researcher, medical doctor and, student or senior PET researcher) was evaluated using Krippendorff’s alpha (α) (Krippendorff, 2011).

SUVr Computation

Each PET image was warped to the MNI-space and normalized using a generic cerebellar GM reference region for $[^{11}\text{C}]\text{PiB}$, and a composite reference region including pons and cerebellar WM for $[^{18}\text{F}]\text{AV45}$ to obtain SUVr images using our in-house pipeline. The mean and maximum

Fig. 1 Transaxial views of $[^{11}\text{C}]\text{PiB}$ (left) and $[^{18}\text{F}]\text{AV45}$ (right) scans, classified as $\text{A}\beta+$, Equivocal, and $\text{A}\beta-$ using both inverse gray-scale and rainbow-scale, displayed as scaled during visual assessment on the clinical viewer with the PET images shown in their native space as in actual clinical reading situation



cortical SUVr values were then measured using generic region of interest (ROI) templates for the cortical regions of frontal, parietal, temporal, occipital, anterior and posterior cingulate, nucleus accumbens, and thalamus (Tanaka et al., 2020).

Automated Binary Classification Algorithms

3 types of automated classification approaches were tested using the two visual assessments separately as ground truth. The first one relied on the determination of global cut-points for mean and maximum SUVr via ROC analysis above which scans are classified as positive. The second approach is based on ML algorithms with regional SUVr values used as features. The third approach used DL convolutional neural networks with 3D PET images as inputs.

Classification Using ROC Analysis and ML Algorithms

ROC analysis as well as 68 different ML models were built and tested for binary classification. A large number of ML were investigated as the ML models differ in their algorithms, implementation and the list of available options. The algorithms were trained starting from 180, 80, and 140 datasets for ADNI- ^{18}F AV45, ADNI- ^{11}C PiB, and CIRC- ^{11}C PiB respectively and progressively decreased in steps of 20. Each time, training and evaluation subjects were randomly selected from the cohort for building up the model and for the evaluation on the unseen data. Each model was fully built and optimized using the selected training data only. During this building process, optimal hyperparameters of the ML method (if any) were determined using the model's default searching grid in a tenfold cross-validation framework. In this process, the training dataset was further partitioned to assess, on slightly different data, each candidate combination of tuning parameters of the model search grid. Across each data set, the performance of the ML method was calculated on the held-out samples and the mean and standard deviation were summarized for each combination. The combination leading to the optimal results are chosen for the final training using the entire training set. The model was then applied to 20 unseen data to determine its performance in terms of classification accuracy, sensitivity, and specificity. The whole process, from random subject selection to parameter optimization and performance evaluation on unseen data was repeated 1000 times to derive the mean and standard deviation of the performance metrics. All computation was done under R (v. 3.6.3) using the caret package (v. 6.0–86, <https://topepo.github.io/caret/available-models.html>).

Classification Using Convolutional Neural Networks with PET Images

Residual network (ResNet) (He et al., 2016) and Squeeze-and-Excitation ResNet (SEResNet) (Hu et al., 2018) were employed as the DL networks in this study. The main feature of ResNet is the use of shortcut connections for identity mapping where outputs from previous layers are added to the outputs of the stacked layers, thus increasing the depths and accuracy of the network. The network won first place in classifying a large dataset of human-annotated photographs in the ImageNet Challenge (ILSVRC) in 2015 (He et al., 2016). SEResNet is built on ResNet with the addition of Squeeze-and-Excitation (SE) blocks, which add parameters to every single convolutional block to enhance the adjustment of the weight for each channel. The input convolutional block is first squeezed into a channel descriptor by average pooling and fed into the activation functions as input-conditioned channel weights, thus introducing dynamics into the network (Hu et al., 2018).

ResNet and SEResNet are available on NiftyNet (Version 0.5.0) (Gibson et al., 2018), a TensorFlow-based convolutional network platform. NiftyNet implements a patch-sampling strategy to extract the necessary information for better convergence and higher performance generalization. The computation was performed on CPU (Dell OptiPlex 9020) and the training progress was tracked using TensorBoard (Version 1.12.2). The networks were kept unchanged for ease of comparison with other works but were each optimized with about 50 different hyperparameter configurations (https://niftynet.readthedocs.io/en/dev/config_spec.html), within the computational feasibility of the CPU and using only CIRC- ^{11}C PiB images. Each time, cross-entropy was employed as the loss function. The final number of iterations of 1000 was subsequently determined from Tensorboard where the loss function curve flattened. Both optimized ResNet and SEResNet employed Adam optimizer with a batch size of 8 and a learning rate of 0.003. However, a spatial window size of (32,32,32) with Parametric Rectified Linear Unit (PReLU) activation function was applied for ResNet while a spatial window size of (24,24,96) with Leaky Relu activation function was applied for SEResNet. A whole head mask was generated by thresholding the PET image intensity and was applied as weighting during network training.

DL networks were trained and tested using ^{18}F AV45 only and the ^{11}C PiB images from both ADNI and CIRC databases pooled together and using either rainbow- or gray-scale binary visual assessment results. The effect of the number of training data on the classification accuracy was further investigated using the better DL network with the minimal variation in the number of training data of

140 and 180, due to the extensive time required to train the networks. The network was trained individually for each visual assessment method 10 times with random selection of the training data and evaluated on the remaining unseen data.

Evaluation of the Performance of Classification Algorithms

The performance of the models for scan classification was evaluated with accuracy (ACC), sensitivity or true positive rate (TPR), and specificity or true negative rate (TNR). Evaluation was carried out independently for each cohort (ADNI-[¹⁸F]AV45, ADNI-[¹¹C]PiB, and CIRC-[¹¹C]PiB) and each visual assessment method (rainbow-scale and gray-scale). Statistical analysis was performed using unpaired 2-tailed t-test using GraphPad (GraphPad Software, CA, US), with significance defined at 0.05, across the different approaches and 2 classification methods (rainbow and gray) for each cohort where possible.

Results

Comparison of Visual Assessment Methods

The confusion matrices resulting from the binary and ternary classifications obtained with the 2 visual assessment

methods and for the 3 cohorts are shown in Table 2. We observed that most discrepancies consisted of subjects who were classified negative with the rainbow-scale but turned positive with the gray-scale. This migration concerned 7.7% (16 out of 209), 3.9% (4 out of 103), and 6.8% (12 out of 176) of the cases for ADNI-[¹⁸F]AV45, ADNI-[¹¹C]PiB, and CIRC-[¹¹C]PiB respectively. Significant differences were observed for binary classifications of ADNI-[¹⁸F]AV45 and CIRC-[¹¹C]PiB only using Wilcoxon Signed-Rank Test, with significance defined at $p < 0.05$.

Generally, high agreement of $\kappa > 0.8$ was found between the two assessment methods for both A β -PET tracers with binary and ternary classification (Table 3 top). Higher agreement was observed for ADNI-[¹¹C]PiB than for the two other cohorts. The inter-rater results showed that the rainbow-scale yielded higher agreement than the gray-scale (Table 3 bottom). The ternary rating was particularly discordant for CIRC-[¹¹C]PiB using the gray-scale, possibly due to the presence of more ambiguous cases caused by concomitant CeVD. Excellent agreements were obtained using rainbow-scale for binary and ternary classifications with ADNI-[¹¹C]PiB data, where about 67% were classified as A β + (Table 2).

Performance Evaluation of Cut-points from ROC Analysis and ML Algorithms with Number of Training Data

Figure 2 shows the classification accuracy of the 3 cohorts as a function of the number of training data, obtained with

Table 2 Percentage (%) distributions in classification using rainbow-scale and gray-scale for ADNI-[¹⁸F]AV45 (top), ADNI-[¹¹C]PiB (middle) and CIRC-[¹¹C]PiB (bottom) with binary (left) and ternary classifications (right)

ADNI-[¹⁸ F]AV45		Rainbow	
		A β -	A β +
Gray	A β -	60.8	0.0
	A β +	7.7	31.6

ADNI-[¹¹ C]PiB		Rainbow	
		A β -	A β +
Gray	A β -	28.2	0.0
	A β +	3.9	68.0

CIRC-[¹¹ C]PiB		Rainbow	
		A β -	A β +
Gray	A β -	57.4	0.0
	A β +	6.8	35.8

ADNI-[¹⁸ F]AV45		Rainbow			
		A β -	A β +	Equivocal-	Equivocal+
Gray	A β -	58.9	0.0	0.0	0.0
	A β +	0.0	27.8	4.3	3.3
	Equivocal-	0.5	0.0	1.4	0.0
	Equivocal+	1.9	0.0	1.4	0.5

ADNI-[¹¹ C]PiB		Rainbow			
		A β -	A β +	Equivocal-	Equivocal+
Gray	A β -	28.2	0.0	0.0	0.0
	A β +	2.9	67.0	1.0	0.0
	Equivocal-	0.0	0.0	0.0	0.0
	Equivocal+	0.0	0.0	0.0	1.0

CIRC-[¹¹ C]PiB		Rainbow			
		A β -	A β +	Equivocal-	Equivocal+
Gray	A β -	52.8	0.0	0.0	0.0
	A β +	2.3	31.3	2.8	3.4
	Equivocal-	4.0	0.0	0.6	0.0
	Equivocal+	0.6	0.0	1.1	1.1

Equivocal cases in ternary classifications were further separated into +/- depending on their final binary classification

Table 3 Agreement in consensus classification between rainbow- and gray-scales with Cohen's Kappa (κ) [confidence interval] and inter-reader agreement among 3 readers with Krippendorff's alpha (α), with 2 and 3 classes classification for [^{11}C]PiB and [^{18}F]AV45

Metrics	Assessment Methods	No of Classes	[^{11}C]PiB			[^{18}F]AV45
			ADNI	CIRC	All	ADNI
Agreement between Gray & Rainbow (κ)		2	0.908 [0.820, 0.996]	0.858 [0.781, 0.935]	0.886 [0.832, 0.940]	0.834 [0.757, 0.911]
		3	0.911 [0.827, 0.996]	0.764 [0.681, 0.847]	0.825 [0.766, 0.885]	0.812 [0.7427, 0.882]
Inter-rater agreement (α)	Rainbow	2	1.000	-	-	0.899 ^c
		3	0.960	-	-	0.730 ^c
	Gray	2	0.867	0.817 ^a	0.845 ^b	0.798
		3	0.501	0.376 ^a	0.422 ^b	0.624

Only consensus data (n = 179) is available for CIRC-[^{11}C]PiB using the rainbow-scale

^aRated by the senior PET researcher instead of the student

^bCIRC-[^{11}C]PiB data was rated by the senior PET researcher while ADNI-[^{11}C]PiB was rated by the student

^cStudent rated only 111 out of 209 scans

cut-points for mean and maximum SUVr determined from ROC analysis, as well as with 4 ML algorithms selected based on their performance. Note that, classification based on mean SUVr cut-points never performed the best.

Classification based on maximum SUVr cut-points in place of mean SUVr lead to increased accuracy for the ADNI-[^{18}F]AV45 cohort with both visual assessment methods and for the CIRC-[^{11}C]PiB cohort with gray-scale only.

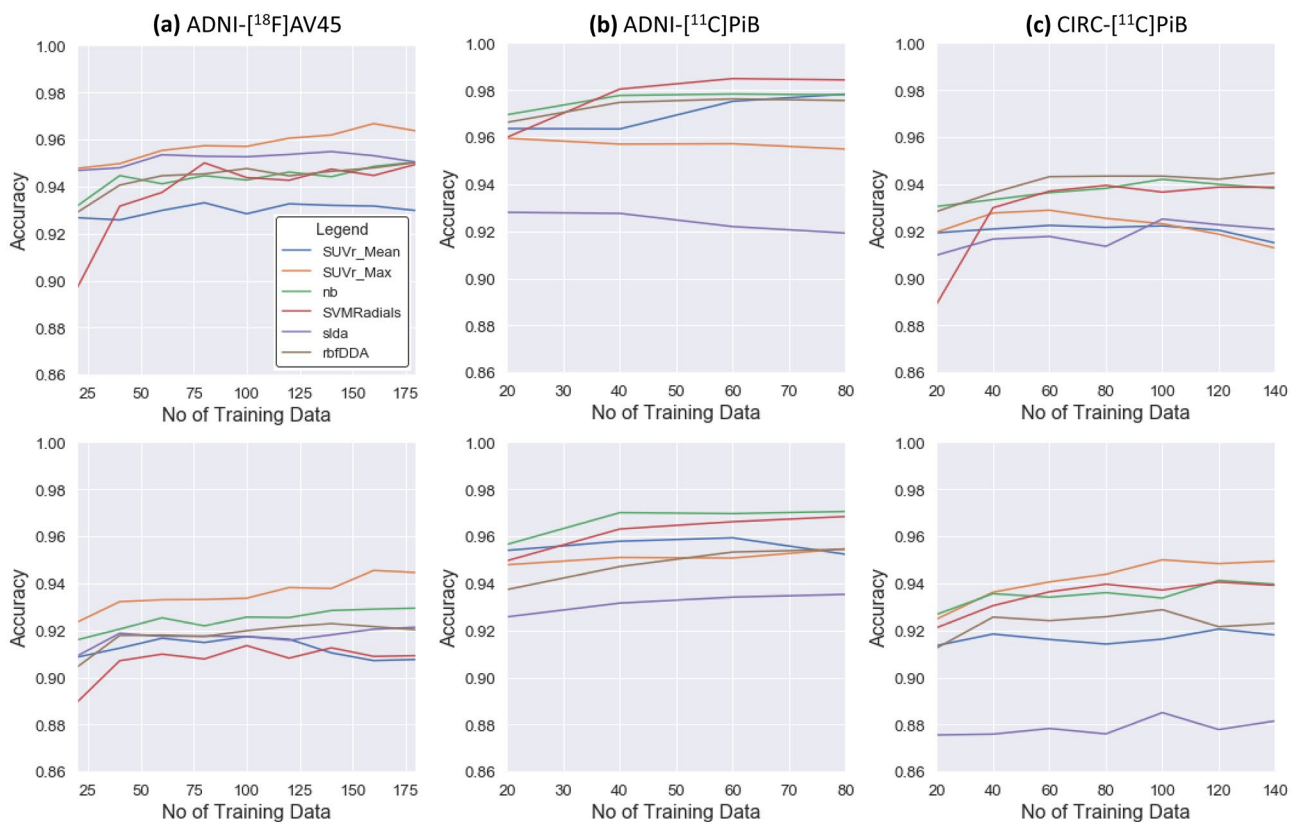


Fig. 2 Classification accuracy of unseen data averaged over the 1000 realizations as a function of the number of training data of (a) ADNI-[^{18}F]AV45, (b) ADNI-[^{11}C]PiB, and (c) CIRC-[^{11}C]PiB, with (top) rainbow-scale and (bottom) gray-scale visual assessment, and obtained with cut-points of mean SUVr (blue), max SUVr (orange)

determined from ROC analysis, and 4 selected ML algorithms: Naïve Bayes (nb, green), Support Vector Machine Radial Weights (SVMRadial, red), Stabilized Linear Discriminant Analysis (slda, purple) and Radial Basis Function Network (rbfDDA, brown)

Supplementary Figures 1 and 2 show that these increases resulted from both higher sensitivity and specificity. However, its performance was below that of mean SUVr in the classification of CIRC- $^{[11]C}$ PiB using rainbow-scale, and ADNI- $^{[11]C}$ PiB using both rainbow- and gray-scales due to reduced specificity. Among the 68 ML algorithms, Naïve Bayes (nb) performed the best overall, specifically with gray-scale assessment. Support Vector Machine Radial Weights (SVMRadial) showed similar performance to that of nb, except for ADNI- $^{[18]F}$ AV45 using gray-scale. Stabilized Linear Discriminant Analysis (slda) performs the best out of the 68 ML algorithms for ADNI- $^{[18]F}$ AV45 using rainbow-scale, but performed the worst for the rest. Radial Basis Function Network (rbfDDA) performed best for CIRC- $^{[11]C}$ PiB with rainbow-scale, but yielded moderate performance for the others. A summary of the performance obtained with the 68 ML algorithms is given in Supplementary Fig. 3 showing that Naïve Bayes classifier was overall the best performing algorithm.

Figure 3 shows the average differences between classification accuracies obtained with training and unseen data as a function of the number of training data for the three cohorts based on the cut-points of mean SUVr (Fig. 3a), and maximum SUVr (Fig. 3b) determined from ROC analysis, and with Naïve Bayes (Fig. 3c). Overall, it shows that Naïve Bayes required less training data to build a classification model that is generalizable to other scans. These results also indicated that the classification performance of cut-points approaches was usually overestimated when reported using training data only. The differences in mean accuracy and

the corresponding statistical differences for the three different methods and two different visual assessment methods employed are shown in Supplementary Tables 1–4. Generally, significant differences were observed across the 3 methods for ADNI- $^{[18]F}$ AV45, regardless of the visual assessment method employed, as well as for both $^{[11]C}$ PiB datasets, trained using the gray-scale visual assessment results.

Performance Evaluation of Deep Learning Approach with Number of Training Data

Although ResNet and SEResNet were evaluated, their performances were quite similar with ResNet performing more consistently and better for all cases. As such, only ResNet was selected for further evaluation with 140 and 180 training data. Table 4 shows the classification performance obtained with ResNet on unseen data when trained with 140 and 180 scans. ADNI- $^{[11]C}$ PiB and CIRC- $^{[11]C}$ PiB were pooled together leading to 2 cohorts. Surprisingly, higher agreement between automatic classification and visual classification was obtained with networks trained using rainbow-scale visual assessment for $^{[18]F}$ AV45 and using gray-scale for $^{[11]C}$ PiB. Higher agreement was also obtained for $^{[11]C}$ PiB, even though the network was trained using two different $^{[11]C}$ PiB datasets. The average classification accuracy generally improved with increasing number of training data except for $^{[18]F}$ AV45 with rainbow-scale. However, the variations in classification accuracy and specificity were generally smaller with rainbow-scale than gray-scale. No statistical difference was observed in the accuracy of ResNet in classification when trained using

Fig. 3 Average differences in accuracy between training and unseen data with the number of training data for 3 classification methods based on (a) mean SUVr, and (b) max SUVr cut-points determined from ROC analysis, and (c) machine learning algorithm of Naïve Bayes (nb) for the 3 datasets of ADNI- $^{[11]C}$ PiB (blue), CIRC- $^{[11]C}$ PiB (gray) and ADNI- $^{[18]F}$ AV45 (orange) with rainbow- (full line) and gray-scale (dashed line) visual assessment classification

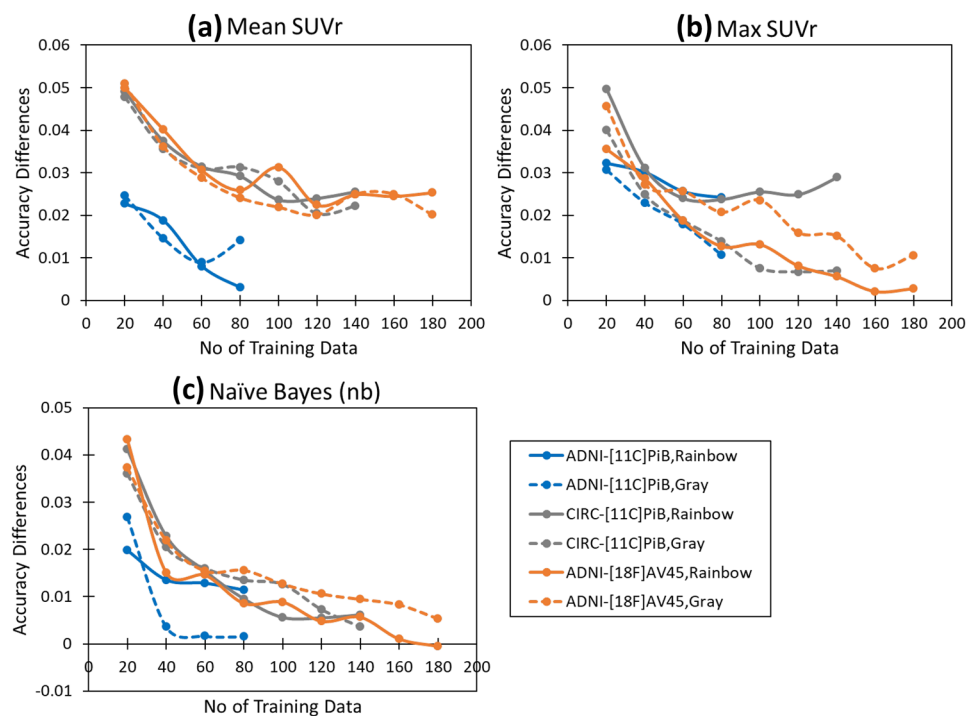


Table 4 Averaged Accuracy (ACC), Sensitivity (TPR), and Specificity (TNR) (mean \pm stdev) in classifying the unseen data of [^{18}F]AV45 and [^{11}C]PiB from both ADNI and CIRC datasets when trained with

140 and 180 training data using rainbow-scale and gray-scale visual assessment classifications

Datasets	No of Training Data	Rainbow-scale			Gray-Scale		
		ACC	TPR	TNR	ACC	TPR	TNR
[^{18}F]AV45	140	0.901 \pm 0.054	0.841 \pm 0.147	0.940 \pm 0.055	0.803 \pm 0.103	0.677 \pm 0.359	0.887 \pm 0.137
	180	0.869 \pm 0.072	0.776 \pm 0.231	0.909 \pm 0.126	0.831 \pm 0.195	0.859 \pm 0.205	0.832 \pm 0.311
[^{11}C]PiB	140	0.889 \pm 0.120	0.831 \pm 0.247	0.953 \pm 0.057	0.904 \pm 0.040	0.859 \pm 0.087	0.958 \pm 0.038
	180	0.922 \pm 0.034	0.893 \pm 0.096	0.947 \pm 0.043	0.953 \pm 0.052	0.893 \pm 0.101	0.912 \pm 0.117

rainbow and gray-scale visual assessment methods (Supplementary Table 4). The distributions of mean and max SUVR values of subjects classified as A β -/+ using ResNet compared to visual assessment classifications using both rainbow- and gray-scales are shown in Supplementary Fig. 4.

Discussion

In this study, we investigated the performance of three approaches for the classification of A β -PET images: (1) global cut-points derived from ROC analysis, (2) ML algorithms using regional SUVR measurements, and (3) a DL network with 3D A β -PET images, under various conditions with the ultimate goal to identify a suitable approach to replace the tedious and consuming visual assessment.

Evaluation of Visual Assessment Methods

Our results contradicted our expectation whereby the recommended method for each radiotracer would be more suitable to interpret the respective scans, with higher inter-reader agreement (Table 3). The results showed that the rainbow-scale seems to be a more reliable assessment procedure for both radiotracers, supported by the higher agreement between raters with different backgrounds and years of experience and for all cohorts (Table 3), and higher consistency between the classification of the scans (AB-, equivocal+, equivocal+, A+) and their SUVR values (Supplementary Fig. 5). The use of a reference region to scale the image intensity likely made the assessment more reproducible than the more subjective intensity scaling performed by adjusting the contrast between WM and GM.

More scans were assessed positive using the gray-scale procedure (Table 2) leading to lower cut-points, but more importantly, this visual assessment procedure yielded higher variability in the mean and maximum SUVR values of the positive group showing important overlaps with the negative group (Supplementary Fig. 5). Many scans that turned positive with the gray-scale assessment exhibited a single focal uptake which did not reach the required intensity with

the rainbow-scale but yielded a loss of contrast between GM-WM with the gray-scale method (Supplementary Fig. 6b). However, the gray-scale allows for quick assessment as there is no need to locate the reference region before adjusting the image intensity. Depending on the reader's experience, the amount of time saved can be as much as 4 times using the gray-scale over the rainbow-scale.

Comparison of Automated Classification Approaches

Generally, the higher the consistency of the ground truth classification (higher agreement between SUVR and visual classification), the higher the performance of the automated classification methods. Their performance differs based on their capacity to deal with focal uptake. Our results first showed that the classical cut-points based approach with mean SUVR was systematically outperformed by the other tested techniques. Classification with maximum SUVR cut-points yielded higher accuracy for ADNI-[^{18}F]AV45 and CIRC-[^{11}C]PiB cohorts due to a higher sensitivity and specificity (Supplementary Figs. 1 and 2). It is intuitively closer to the visual assessment process in particular for the detection of focal uptake which is more common in the ADNI-[^{18}F]AV45 and CIRC-[^{11}C]PiB cohorts. For example, one subject scanned with [^{11}C]PiB was classified as A β - using rainbow-scale but was A β + using gray-scale due to focal regional uptake (Supplementary Fig. 6b). In this case, the maximum SUVR value within the focal region was 2.16, but the global mean SUVR was only about 1.10 due to lack of specific uptake in the remaining cortical GM.

Among the 68 tested ML algorithms, Naïve Bayes performed the best overall, specifically with gray-scale classification. Comparing classification based on global cut-points from ROC analysis and ML algorithms using regional values, ML-based classification achieved better convergence between training and unseen data, and with a smaller number of training data (Fig. 3). Previous attempts to classify A β -PET images automatically using ML algorithms namely SVM with linear kernel or histogram of oriented 3D gradients, achieved high accuracy of > 96%, but with the exclusion of equivocal cases (Cattell et al., 2015; Vandenberghe

et al., 2013). This situation is close to our ADNI- ^{11}C PiB cohort, with few equivocal cases, and for which accuracy above 98% was obtained with some ML methods. However, only a small number of ML algorithms performed better than traditional cut-points (Supplementary Fig. 3), with Naïve Bayes performing the best overall. Although ML algorithm can be a better quantitative approach to support A β -PET image classification than traditional SUVr cut-points, careful selection and validation of ML algorithm are required before clinical use.

The deep learning VGG16 model implemented by Kang et al. yielded higher accuracy of > 92% using image slices compared to 3D volumes of ^{18}F Florbetaben images, with accuracy of > 89%, with either whole brain or GM masks input (Kang et al., 2018). Son et al. (2020) also used 2D and 3D deep learning models to classify ^{18}F Florbetaben images and obtained 100% accuracy for definite A β -/+ cases but with 31.5% discordance in equivocal cases. We obtained comparable high accuracy of about 90% and 95% for unseen data of ^{18}F AV45 and ^{11}C PiB datasets (Table 4). Higher variation in accuracy was obtained with gray-scale, particularly for ^{18}F AV45, indicating that the DL network could not map the links between the visual assessment classification and the ^{18}F AV45 images with 16 subjects that were classified differently. However, the networks were still able to classify the ^{11}C PiB images despite the same number of subjects being classified differently (Table 2). The best-trained network, using rainbow-scale visual assessment results, classified most subjects correctly except for a few cases near the cut-points for both ^{11}C PiB and ^{18}F AV45 (Supplementary Fig. 4).

Methodological Considerations and Study Limitations

Small sample size, particularly for ADNI- ^{11}C PiB, was used. Moreover, the ADNI- ^{11}C PiB dataset consisted of mostly A β + scans, while CIRC- ^{11}C PiB consisted of more A β - and equivocal cases (Table 1). This helps to balance out the dataset. Whole head mask was used in our study unlike other studies, which used either the whole brain mask or gray matter mask, derived from MR images. We chose to use whole head mask as we want to investigate the feasibility to classify the images automatically assuming the subjects only acquired PET scans. This might be the reason for the poorer binary classification results obtained with DL approach compared to that obtained by Son et al. (2020). The optimized DL network was trained with ^{11}C PiB PET images from both ADNI and CIRC datasets rather than trained for individual datasets. This may show that DL networks require more images or a more balanced dataset in order to classify the images with higher accuracy, which may be more important than

the image quality. DL networks were only trained 10 times compared to ROC and ML approaches due to the much longer times required. As such, evaluation was carried out on all remaining unseen data instead of 20 unseen data at each iteration.

In this work, hyperparameters of each ML method were determined automatically in a cross-validation framework to address the time consuming manual optimization task. As a matter of fact, manually optimizing the parameters of each ML model and for each combination of training and evaluation datasets would have been practically impossible. However, we occasionally manually optimized some parameters and compared with the automatically selected parameters, a verification exercise which in the vast majority of the cases confirmed the reliability of the automated procedure. Different reference regions were selected for SUVr computations of ^{11}C PiB and ^{18}F AV45 and this may impact ML classification accuracy. However, assessing this impact is beyond the scope of this study. We do not have the absolute ground truth based on pathology, thus clinical visual assessment was used as ground truth, similar to that used by Son et al. (2020). We used two visual assessment methods to generate the ground truths to determine the impact on three automated classification methods. Our work showed that the classification results and reader agreement varied depending on the visual assessment methods used. The inclusion of more accurate quantification, such as binding potentials via kinetic modelling, would help in checking the visual assessment classification. However, only static PET images were available, hence we were unable to include in this study. Absolute ground truth might be obtained using synthetic images, such as obtained via Monte Carlo simulation, but this comes with other challenges that are beyond the scope of this work.

Conclusions

Higher accuracy was generally obtained for all three approaches when trained with rainbow-scale classification. Automated algorithms cannot replace visual assessment completely as they could not detect focal uptake with 100% accuracy. However, they yielded more consistent results than across raters and thus can be used to support readers in classifying the images. Only a small number of ML algorithms performed better than traditional cut-points, with Naïve Bayes performing the best overall. However, ML-based classification provided the highest reliability even with a small number of training data, with similar accuracy to ROC classification. This shows the promising use of ML algorithms in supporting A β -PET image classification than traditional SUVr cut-points, but careful selection and validation of ML algorithms are

required. Although image-based classification using ML and DL approaches can be achieved without tracer-specific target regions and cut-points, our results showed that DL networks can support the classification of definite cases accurately, but adds very little value as they are also obvious for readers to classify visually. However, they may add as an additional check for equivocal cases, on top of semi-quantitative metrics.

Information Sharing Statement

The datasets generated during and/or analyzed during the current study are not publicly available as the authors have no permission to share the data but are available from the corresponding author on reasonable request.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12021-022-09587-2>.

Acknowledgements We acknowledge all the coordinators from NUH memory clinic and Memory Aging and Cognition Centre for their contributions in subject recruitment and data acquisition, and the PET radiochemistry team from CIRC for the production of [¹¹C]PiB used in this study. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California

Author Contributions Ying-Hwey Nai and Anthonin Reilhac designed and supervised the study. Yee-Hsin Tay, Tomotaka Tanaka, Anthonin Reilhac and Ying-Hwey Nai carried out the visual assessments. Yee-Hsin Tay and Ying-Hwey Nai and Anthonin Reilhac analyzed the data and wrote the draft manuscript. Edward G. Robins was responsible for the production and delivery of [¹¹C]PiB used in the study. Christopher P. Chen designed the clinical study and is responsible for clinical data. All authors contributed to the revision of the draft manuscript and approved the final version of the manuscript for submission.

Funding This study was supported by the following National Medical Research Council grant in Singapore: NMRC/CG/NUHS/2010-R-184-005-184-511, NMRC/CG/013/2013, and NMRC/CIRG/1446/2016.

Availability of Data and Material The data will not be available as the authors have no permission to share the data.

Declarations

Informed Consent Informed consent was obtained from all individual participants included in the study. Patients signed informed consent regarding publishing their data and photographs.

Conflicts of Interest The authors declare that they have no conflict of interest.

Research Involving Human Participants All procedures performed in 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. Ethics approval was obtained from the National-Healthcare Group Domain-Specific Review Board in Singapore.





References

- Cattell, L., Platsch, G., Pfeiffer, R., Declerck, J., Schnabel, J. A., & Hutton, C. (2015). Classification of amyloid status using machine learning with histograms of oriented 3D gradients. *NeuroImage Clinical*, 12, 990–1003. <https://doi.org/10.1016/j.nicl.2016.05.004>
- Eli Lilly. (2012). *Highlights of prescribing information Amyvid (florbetapir F 18 injection)*. Revised December 2019 from <https://pi.lilly.com/us/amyvid-uspi.pdf>
- Gibson, E., Li, W., Sudre, C., et al. (2018). NiftyNet: A deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 158, 113–122. <https://doi.org/10.1016/j.cmpb.2018.01.025>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).
- Jack, C. R., Bennett, D. A., Blennow, K., et al. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4), 535–562. <https://doi.org/10.1016/j.jalz.2018.02.018>
- Johnson, K. A., Minoshima, S., Bohnen, N. I., et al. (2013). Appropriate use criteria for Amyloid PET: A report of the Amyloid imaging task force for the society of nuclear medicine and molecular imaging and the Alzheimer's association. *Journal of Nuclear Medicine*, 54(3), 476–490. <https://doi.org/10.2967/jnumed.113.120618>
- Kang, H., Kim, W. -G., Yang, G. -S., et al. (2018). VGG-based BAPL Score Classification of 18F-Florbetaben Amyloid Brain PET. *Bio-medical Science Letters*, 24(4), 418–425. <https://doi.org/10.15616/bsl.2018.24.4.418>
- Kim, J. P., Kim, J., Kim, Y., et al. (2020). Staging and quantification of florbetaben PET images using machine learning: Impact of

- predicted regional cortical tracer uptake and amyloid stage on clinical outcomes. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(8), 1971–1983. <https://doi.org/10.1007/s00259-019-04663-3>
- Krippendorff, K. (2011) *Computing Krippendorff's Alpha-Reliability* (p. 12). Dep Pap. http://repository.upenn.edu/asc_papers
- Lundeen, T. F., Seibyl, J. P., Covington, M. F., Eshghi, N., & Kuo, P. H. (2018). Signs and Artifacts in Amyloid PET. *Radiographics*, 38(7), 2123–2133. <https://doi.org/10.1148/rg.2018180160>
- McHugh, M. L. (2012) Interrater reliability: The kappa statistic. *Biochemistry Medica*, 22(3), 276–282. <https://doi.org/10.11613/bm.2012.031>
- Ng, S., Villemagne, V. L., Berlangieri, S., et al. (2007). Visual assessment versus quantitative assessment of 11C-PIB PET and 18F-FDG PET for detection of Alzheimer's disease. *Journal of Nuclear Medicine*, 48(4), 547–552. <https://doi.org/10.2967/jnumed.106.037762>
- Reilhac, A., Merida, I., Irace, Z., et al. (2018). Development of a dedicated rebinner with rigid motion correction for the mMR PET/MR Scanner, and Validation in a Large Cohort of 11C-PIB Scans. *Journal of Nuclear Medicine*, 59(11), 1761–1767. <https://doi.org/10.2967/jnumed.117.206375>
- Rowe, C. C., & Villemagne, V. L. (2013). Brain amyloid imaging. *Journal of Nuclear Medicine Technology*, 41(1), 11–18. <https://doi.org/10.2967/jnumed.110.076315>
- Son, H. J., Oh, J. S., Oh, M., et al. (2020). The clinical feasibility of deep learning-based classification of amyloid PET images in visually equivocal cases. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(2), 332–341. <https://doi.org/10.1007/s00259-019-04595-y>
- Tanaka, T., Stephenson, M. C., Nai, Y. H., et al. (2020). Improved quantification of amyloid burden and associated biomarker cut-off points: Results from the first amyloid Singaporean cohort with overlapping cerebrovascular disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 47(2), 319–331. <https://doi.org/10.1007/s00259-019-04642-8>
- Vandenberghe, R., Nelissen, N., Salmon, E., et al. (2013). Binary classification of 18F-flutemetamol PET using machine learning: Comparison with visual reads and structural MRI. *NeuroImage*, 64(1), 517–525. <https://doi.org/10.1016/j.neuroimage.2012.09.015>
- Yamane, T., Ishii, K., Sakata, M., et al. (2017). Inter-rater variability of visual interpretation and comparison with quantitative evaluation of 11C-PiB PET amyloid images of the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) multicenter study. *European Journal of Nuclear Medicine and Molecular Imaging*, 44(5), 850–857. <https://doi.org/10.1007/s00259-016-3591-2>

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

Authors and Affiliations

Ying-Hwey Nai¹  · Yee-Hsin Tay² · Tomotaka Tanaka^{3,4} · Christopher P. Chen^{4,5}  · Edward G. Robins^{1,6}  · Anthonin Reilhac¹  · for the Alzheimer's Disease Neuroimaging Initiative

¹ Clinical Imaging Research Centre, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

² Nanyang Junior College, Singapore, Singapore

³ Department of Neurology, National Cerebral and Cardiovascular Center, Osaka, Japan

⁴ Memory Aging and Cognition Centre, Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

⁵ Memory Aging and Cognition Centre, Department of Psychological Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

⁶ Singapore BioImaging Consortium (SBIC), Agency for Science, Technology and Research (A*Star), Singapore, Singapore