European Radiology (2022) 32:6965–6976
https://doi.org/10.1007/s00330-022-09081-y

HEAD AND NECK

How segmentation methods affect hippocampal radiomic feature accuracy in Alzheimer's disease analysis?

Qiang Zheng[1] · Yiyu Zhang[1] · Honglun Li[2] · Xiangrong Tong[1] · Minhui Ouyang[3]

Received: 9 March 2022 / Revised: 30 June 2022 / Accepted: 3 August 2022 / Published online: 24 August 2022

Abstract

Objectives Hippocampal radiomic features (HRFs) can serve as biomarkers in Alzheimer's disease (AD). However, how different hippocampal segmentation methods affect HRFs in AD is still unknown. The aim of the study was to investigate how different segmentation methods affect HRF accuracy in AD analysis.

Methods A total of 1650 subjects were identified from the Alzheimer's Disease Neuroimaging Initiative database (ADNI). The mini-mental state examination (MMSE) and Alzheimer's disease assessment scale (ADAS-cog13) were also adopted. After calculating the HRFs of intensity, shape, and textural features from each side of the hippocampus in structural magnetic resonance imaging (sMRI), the consistency of HRFs calculated from 7 different hippocampal segmentation methods was validated, and the performance of machine learning–based classification of AD vs. normal control (NC) adopting the different HRFs was also examined. Additional 571 subjects from the European DTI Study on Dementia database (EDSD) were to validate the consistency of results.

Results Between different segmentations, HRFs showed a high measurement consistency ($R > 0.7$), a high significant consistency between NC, mild cognitive impairment (MCI), and AD (T-value plot, $R > 0.8$), and consistent significant correlations between HRFs and MMSE/ADAS-cog13 ($p < 0.05$). The best NC vs. AD classification was obtained when the hippocampus was sufficiently segmented by primitive majority voting (threshold = 0.2). High consistent results were reproduced from independent EDSD cohort.

Conclusions HRFs exhibited high consistency across different hippocampal segmentation methods, and the best performance in AD classification was obtained when HRFs were extracted by the naïve majority voting method with a more sufficient segmentation and relatively low hippocampus segmentation accuracy.

Key Points

- *The hippocampal radiomic features exhibited high measurement/statistical/clinical consistency across different hippocampal segmentation methods.*
- *The best performance in AD classification was obtained when hippocampal radiomics were extracted by the naïve majority voting method with a more sufficient segmentation and relatively low hippocampus segmentation accuracy.*

Keywords Radiomic features · Hippocampus segmentation · Alzheimer's disease · Machine learning · Magnetic resonance imaging

✉ Qiang Zheng
zhengqiang@ytu.edu.cn

[1] School of Computer and Control Engineering, Yantai University, No30, Qingquan Road, Laishan District, Yantai 264005, Shandong, China

[2] Departments of Medical Oncology and Radiology, Affiliated Yantai Yuhuangding Hospital of Qingdao University Medical College, Yantai 264000, China

[3] Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Abbreviations

ACC	Accuracy
AD	Alzheimer's disease
ADAS-cog13	Alzheimer's disease assessment scale
ADNI	The Alzheimer's Disease Neuroimaging Initiative database
ANTs	The Advanced Normalization Tools
AUC	The area under the ROC curve
EDSD	The European DTI Study on Dementia database

HRFs	Hippocampal radiomic features
ICC	Intraclass correlation coefficient
LLL	Local label learning
MAIS	Multi-atlas image segmentation
MCI	Mild cognitive impairment
ML	Metric learning
MMSE	Mini-mental state examination
MV	Majority voting
NC	Normal control
NLP	Nonlocal patch
RF	Random forest
RF-SSLP	Random forest-semi-supervised label propagation
RLBP	Random local binary pattern
ROC	Receiver operating characteristic
SEN	Sensitivity
SPE	Specificity
SVM	Support vector machine

Introduction

Alzheimer's disease (AD) is one of the most common causes of dementia in elder individuals and is a fatal neurodegenerative disease characterized by progressive cognition impairment [1–3], where mild cognitive impairment (MCI) is usually considered a transitional stage between normal aging and early dementia [4, 5]. Radiomics has been demonstrated as a powerful method to extract comprehensive information from specific medical image regions [6–10], including intensity, shape, and texture features. Since hippocampal morphological change is one of the main hallmarks of AD/MCI, hippocampal radiomic features (HRFs) have been used as robust neuroimaging biomarkers for clinical application in AD/MCI based on a multisite structural magnetic resonance imaging (sMRI) study [8–11]. However, how different hippocampal segmentation methods affect HRFs when applied in AD analysis is still unknown.

The prerequisite of HRFs calculation is hippocampal segmentation from MR images, and a variety of hippocampal segmentation methods have emerged at present. In the previous study, we compared the performances from different hippocampal segmentation methods, including majority voting (MV) [12], nonlocal patch (NLP) [13], random local binary pattern (RLBP) [14], metric learning (ML) [15], local label learning (LLL) [16], random forest (RF) [17], and random forest-semi-supervised label propagation (RF-SSLP) [18] using the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). Interestingly, the machine learning–based multi-atlas image segmentation (MAIS) methods have high computational cost, but the segmentation accuracy is only slightly improved [18]. One issue is raised regarding how much the difference of segmentation accuracy affects HRFs and therefore the classification between AD and normal control (NC) or between AD and MCI. The aim of the study is to investigative how segmentation methods affect HRF accuracy in AD analysis, and find out the optimal hippocampus segmentation method and the associated HRFs that achieve the best performance of classification between AD and NC or between AD and MCI. In our study, the different segmentation methods under comparison include MV [12], NLP [13], RLBP [14], ML [15], LLL [16], RF [17], and RF-SSLP [18]. For a comprehensive comparison, a series of threshold values of 0.1–0.9 with a step size of 0.1 was adopted in the MV method to obtain the sufficient segmentation or under segmentation. We hypothesize that different segmentation methods will not significantly affect HRFs when used in AD analysis, and the hippocampal surrounding area covering voxels with small gray matter volume at the edge of the hippocampus obtained by a more sufficient segmentation can also contribute valuable information when extracting radiomic features for AD analysis.

Materials and methods

Data acquisition and processing

In our study, 1650 participants consisting of 603 NC, 764 MCI, and 283 AD subjects were obtained from the ADNI dataset (http://adni.loni.usc.edu). The clinical measurements of mini-mental state examination (MMSE) score and Alzheimer's disease assessment scale (ADAS-cog13) [19] for those subjects were also identified from the ADNI cohort (https://ida.loni.usc.edu/pages/access/studyData.jsp) and are summarized in Table 1. The ADNI study was approved by the institutional review boards of all the participating institutions. Informed written consent was obtained from all participants across ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 studies [20].

An additional dataset of 571 subjects consisting of 230 NC, 183 MCI, and 158 AD subjects was obtained from the EDSD cohort (The European DTI Study on Dementia, https://neugrid4you.eu) to serve as a validation dataset in the present study. The demographic characteristics and MMSE

Table 1 The detailed information about the subjects from the ADNI cohort

	NC (N = 603)	MCI (N = 764)	AD (N = 283)	p
Age (years)	73.46 ± 6.17	72.96 ± 7.70	74.91 ± 7.70	< 0.001
Gender (M/F)	277/326	447/317	152/131	< 0.001
MMSE	29.08 ± 1.10	27.57 ± 1.81	23.18 ± 2.14	< 0.001
ADAS-cog13	10.37 ± 4.37	16.63 ± 6.67	30.03 ± 7.91	< 0.001

score for those subjects were also identified from the EDSD cohort (details can be found in the supplementary materials S01). The dataset from EDSD served as an independent and supplementary dataset to validate the consistency of the result obtained from ADNI.

The hippocampus segmentation

For each subject, the T1-weighted MR image was aligned to Montreal Neurological Institute (MNI) space using a linear registration method and resampled to 1 × 1 × 1 mm^3 after N4 correction using the Advanced Normalization Tools (ANTs) (https://github.com/ANTsX/ANTs), followed by identifying bounding boxes that were large enough to cover the hippocampal region in the MNI space. The bilateral hippocampus was then segmented using different segmentation methods, including MV [12], NLP [13], RLBP [14], ML [15], LLL [16], RF [17], and RF-SSLP [18]. Detailed information about those segmentation methods can be found in our previous study [18].

Besides, since most existing MAIS methods were chasing hippocampus segmentation accuracy, the under segmentation or sufficient segmentation of the hippocampus was generated for a comprehensive comparison when using HRFs in classification between AD and NC or between AD and MCI. The under or sufficient segmentation was produced by segmenting bilateral hippocampus by the MV method with a series of threshold values from 0.1 to 0.9 (step size = 0.1) [12] (Fig. 1a). The visualization of hippocampal segmentation obtained from different segmentation methods for the sampled participants is provided in supplementary materials S02.

Hippocampal radiomic feature calculation

Based on the hippocampus segmentation achieved above, the radiomic features of the bilateral hippocampus were computed by a publicly available MATLAB script (https://github.com/YongLiulab) [10]. Specifically, a total of 55 features (14 intensity features, 8 shape features, and 33 textural features) were calculated for each side hippocampus, resulting in 55 * 2 = 110 features for each individual (including left and right hippocampus). The intensity features were the first-order statistical distribution of voxel intensities of the hippocampus, the shape features were the descriptors of the 3D size and shape of the hippocampus, and the textural features were describing the spatial distribution of voxel intensities of the hippocampus. The definition of each radiomic feature and the parameters for calculation can be found in the previous studies [6, 10] (Fig. 1b) and supplementary materials S03.

Statistical analysis

The statistical analysis was designed to validate the consistency or reveal the difference of HRFs in AD analysis based on different segmentation methods, including

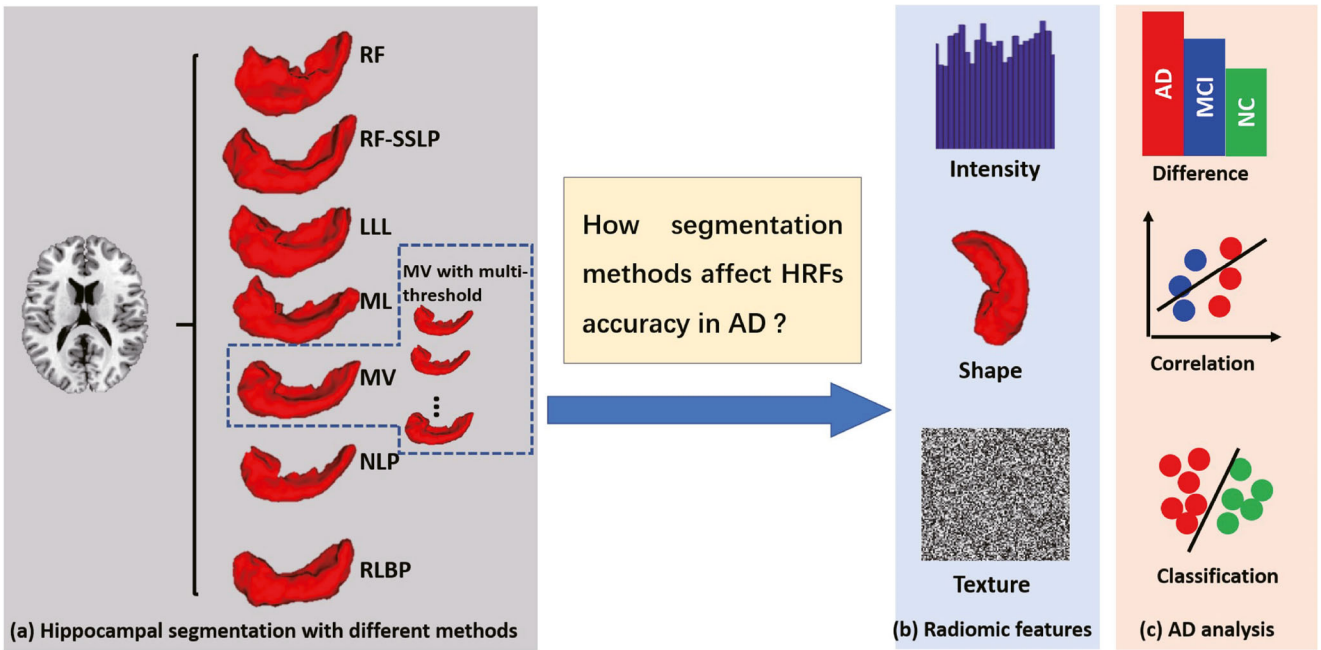

Fig. 1 Schematic of the data analysis pipeline. The hippocampus was segmented using different hippocampal segmentation methods in **a**; radiomic features of each hippocampal segmentation were extracted in **b**; performing the group difference, correlation, and classification analysis in AD, MCI, and NC in **c**. NC, normal control; MCI, mild cognitive impairment; AD, Alzheimer's disease

measurement consistency, statistical consistency, and clinical consistency. The present study consisted of three experiments described below after normalizing each HRF among individuals using a max-min standardized method (Fig. 1c). The Bonferroni correction was adopted to correct the multiple comparisons.

To examine the measurement consistency of HRFs obtained from different segmentation methods, the Pearson correlation of HRFs between all unique paired segmentation methods was performed for each individual. In our study, 7 segmentation methods of MV, NLP, RLBP, ML, LLL, RF, and RF-SSLP were adopted, and the number of unique pairs was $7 \times 6/2 = 21$. For each unique paired segmentation method, 1650 *R*-values corresponding to all subjects in this study were obtained. Furthermore, the distribution of *R*-values was delineated to evaluate whether the high consistency could be obtained for most subjects in each pair of segmentation methods.

To evaluate the statistically significant consistency of HRFs in the MCI and AD groups, three group comparisons of AD vs. NC, MCI vs. NC, and AD vs. MCI were performed on HRFs by the two-sample two-sided *T*-test, in which, each group comparison was implemented upon the 7 segmentation methods. The statistical significance of each HRF quantified by a series of *T*-values was then evaluated after removing the age and sex effects with a linear regression model (i.e., Radiomic feature = Original radiomic feature − ($W_1 \times$ age + $W_2 \times$ sex)). Then, the statistically significant consistency of HRFs in the MCI and AD groups was determined by the Pearson correlation of *T*-values between all unique paired segmentation methods in each group comparison.

The clinical consistency was defined as the correlation between the radiomic features and clinical measures across different hippocampal segmentation methods. To further validate the clinical consistency of HRFs under different segmentation methods, the Pearson correlation was performed between HRFs and clinical measurements of MMSE and ADAS-cog13 in the MCI and AD groups. The clinical consistency of HRFs in the MCI and AD groups was determined by an additional Pearson correlation between the *R*-values of all unique paired segmentation methods.

Furthermore, to evaluate whether the different MRI scanning protocols would significantly affect the consistency of HRFs across different segmentation methods, three subsets with different parameters were identified from all involved ADNI subjects for conducting a re-analysis. Of the original 1650 subjects in ADNI, 1275 participants were found with the same 1.2-mm slice thickness but with 1.5-T/3-T field strength, and the remaining 375 participants without clarified field strength were excluded. Therefore, group 1 (N = 506, 172NC, 262MCI, and 72AD) was at 3-T field strength with 1.2-mm slice thickness, group 2 (N = 769, 216NC, 374MCI, and 179AD) was at 1.5-T field strength with 1.2-mm slice thickness, and group 3 (N = 1275, 388NC, 636MCI, and 251AD) was the combination of group 1 and group 2.

Machine learning–based classification analysis

Other than the statistical analysis above, the machine learning–based classification of NC and AD using different HRFs obtained from the 7 segmentation methods was also carried out. Specifically, a nonlinear support vector machine (SVM) model with a radial basis function kernel was adopted based on the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) and the classification performance was determined by calculating the accuracy (ACC), specificity (SPE), sensitivity (SEN), and area under the receiver operating characteristic (ROC) curve (AUC) with a 10-fold cross-validation strategy (Fig. 1c).

To further expand the segmentation accuracy and obtain more comprehensive segmentations of sufficient segmentation and under segmentation, the MV method implemented by a series of threshold values from 0.1 to 0.9 (step size = 0.1) was also carried out, followed by the machine learning–based classification analysis above. Typically, threshold = 0.5 was used in a number of studies [12], threshold < 0.5 indicates a sufficient-segmentation where the MV result covers more surrounding voxels with small gray matter volume at the edge of the hippocampus, and threshold > 0.5 indicates an under segmentation.

In addition, the classification of AD and MCI under the same setting above was also conducted based on the 7 segmentation methods. Specifically, due to the high imbalance of AD subjects and MCI subjects in ADNI, the MCI with equal amounts of AD (N = 283) was first randomly sampled for 100 times. After that, for each time, the 10-fold cross-validation was performed using the same machine learning–based classification method as AD vs. NC. At last, the performance of HRFs under different segmentation methods in classifying AD and MCI was assessed by the mean measures of 100 10-fold cross-validation.

Test-retest analysis to identify the most reliable radiomic features

Given the high consistency of the radiomic features between all segmentation methods, a test-retest analysis for each radiomic feature was performed to identify the most or least reliable radiomic features. Briefly, intraclass correlation coefficient (ICC; ICC = (BMS − WMS)/BMS) was calculated to estimate the reliability in different segmentation methods when measuring each radiomic feature, where BMS is the between-subjects mean square, and WMS is the within-subject mean square [21, 22]. In the present study, the radiomic features with ICC > 0.7 were with high reliability [22].

Results

Demographic characteristics and neuropsychological assessments

The 1650 subjects (ADNI), including 603 NCs, 764 MCI, and 283 AD subjects, were identified in the present study. Among the three groups, significant differences were observed in age ($p < 0.001$, ANOVA test) and sex ($p < 0.001$, chi-square test). Besides, the clinical measures (MMSE score and ADAS-cog13 score) were significantly different among the three groups ($p < 0.001$, ANOVA test) (Table 1).

A total of 571 subjects (EDSD), including 230 NCs, 183 MCI, and 158 AD subjects, were also identified in the present study. Among the three groups, significant differences were observed in age ($p < 0.001$, ANOVA test) and sex ($p < 0.001$, chi-square test). Besides, the MMSE score was significantly different among the three groups ($p < 0.001$, ANOVA test) (Table S1 in supplementary materials S01).

High consistency of radiomics features obtained from different segmentation methods

(1) Measurement consistency: For most subjects (55–84% under different paired methods, mean 72%, std 8.5), the HRFs showed a high consistency between all unique paired segmentation methods, where the R-value of the Pearson correlation was bigger than 0.7 (Fig. 2). The highest consistency of the HRFs (more than 80% of subjects) happened between RF and RLBP, RF-SSLP and

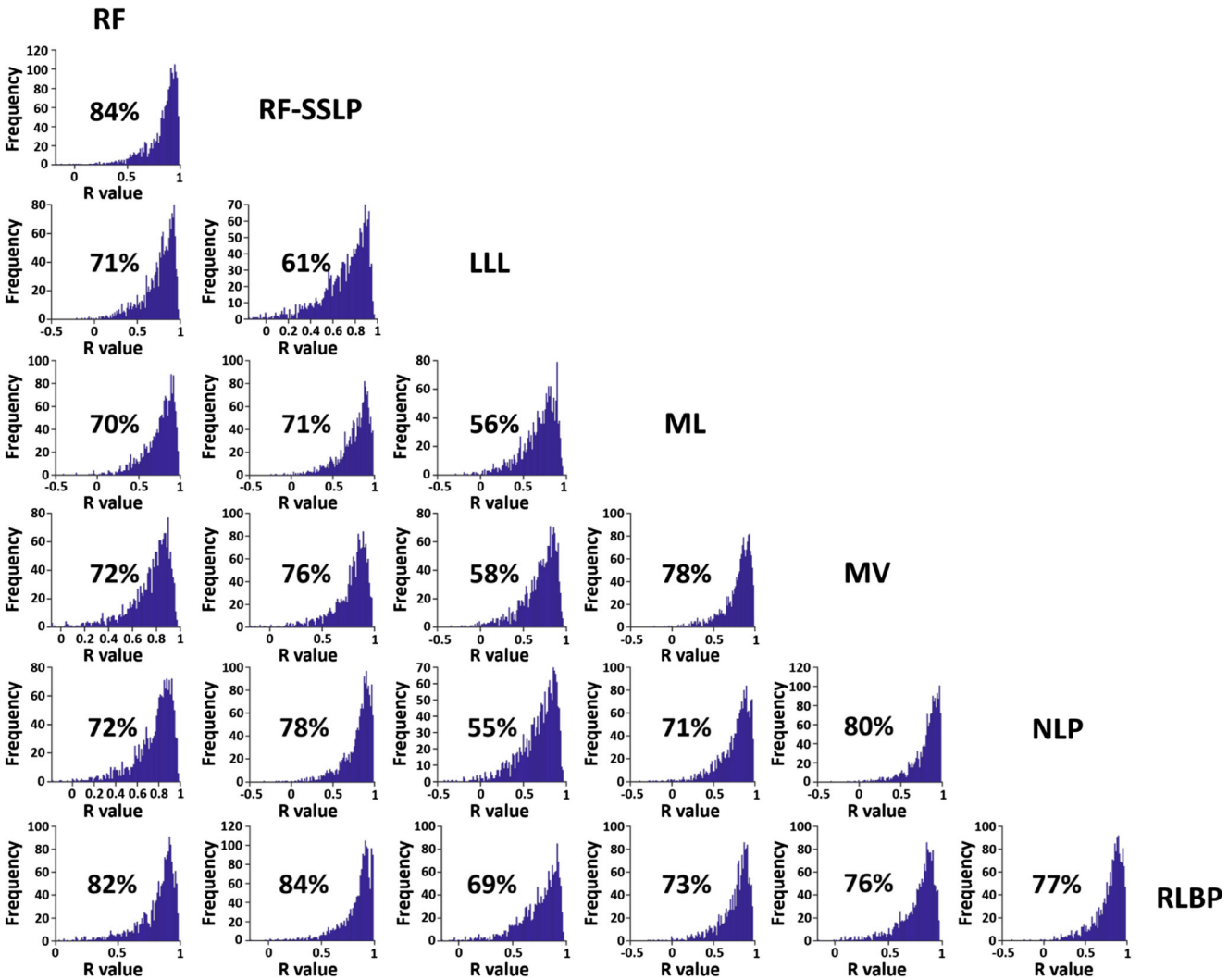

Fig. 2 The distribution of the *R*-value of correlation between the radiomic features was computed with seven different hippocampal segmentation methods. The number indicates how much the ratio is with *R*-value > 0.7

RLBP, and RF and RF-SSLP hippocampal segmentation methods (with the R-value > 0.7).

(2) Statistical consistency: When comparing groups NC and AD (Fig. 3a), the HRFs showed a high statistically significant consistency despite different segmentation methods according to the color observation of the T-value plot. Of the 110 HRFs involved, significant differences were observed in a variety of HRFs between AD and NC ($p < 0.05$), especially in the "Size," "Area," "Compactness," "GLN," and "RLN" ($p < 10^{-10}$). High R-values of the Pearson correlation of the T-values were also observed between all unique paired segmentation methods (R-value > 0.8, Fig. 3d). Additionally, the same consistency was also observed in group comparison of NC vs. MCI (Fig. 3b), and MCI vs. AD (Fig. 3c), also with R-value > 0.8 between all unique paired segmentation methods (Fig. 3e and f).

(3) Clinical consistency: The significant correlation was observed between the HRFs and MMSE score in MCI and AD groups with $p < 0.05$, especially in the "Size," "Area," "Compactness," "GLN," and "RLN" ($p < 10^{-10}$). More importantly, the results showed that the correlation values were significantly correlated between all unique paired segmentation methods (R-value > 0.9) (Fig. 4a and c). Besides, the significant correlations between the HRFs and ADAS-cog13 score in MCI and AD group were also found in this study, and the associated correlation values were significantly correlated between all unique paired segmentation methods (R-value > 0.9) (Fig. 4b and d).

(4) The re-analysis between subsets of ADNI: The re-analysis demonstrated that the results of group 1, group 2, and group 3 subsets were of great similarity with the whole ADNI database in terms of measurement/statistical/clinical consistency of HRFs across different hippocampal segmentation methods, highlighting the reproducibility of consistency of HRFs under different MRI scanning protocols (detailed results can be found in supplementary materials S04). Moreover, ADNI is a large complicated database comprising more than 50 sites

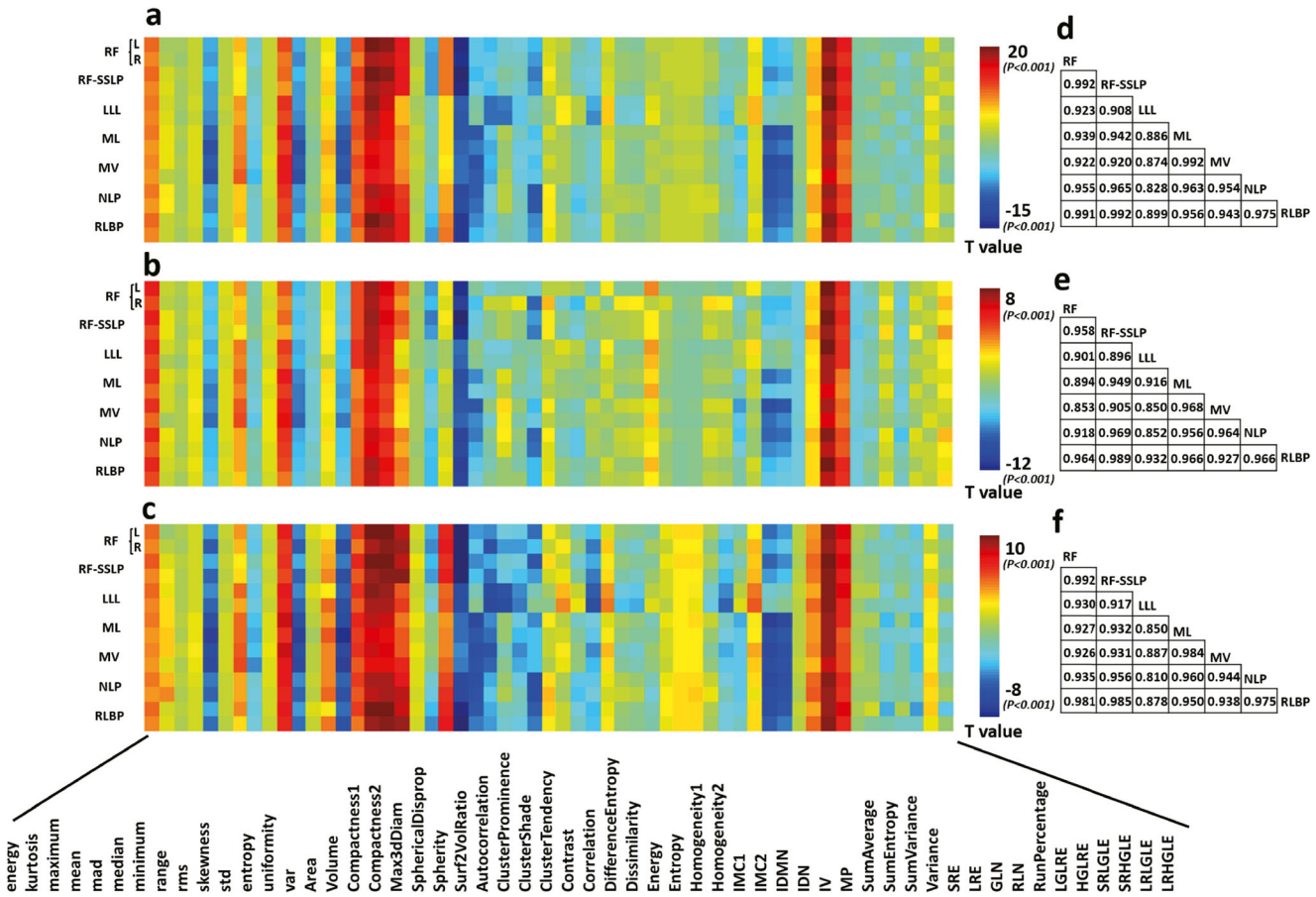

Fig. 3 The difference map of the radiomics features. **a**–**c** The *T*-values of the statistical difference of radiomic features between AD and NC in **a**, between NC and MCI in **b**, and between MCI and AD in **c** were calculated based on seven hippocampal segmentation methods. **d**–**f** The *R*-values of correlation between the *T*-values of the above difference analysis in **a**–**c** respectively were calculated between the seven hippocampal segmentation methods. L means left and R means right of the hippocampus. NC, normal control; AD, Alzheimer's disease; MCI, mild cognitive impairment

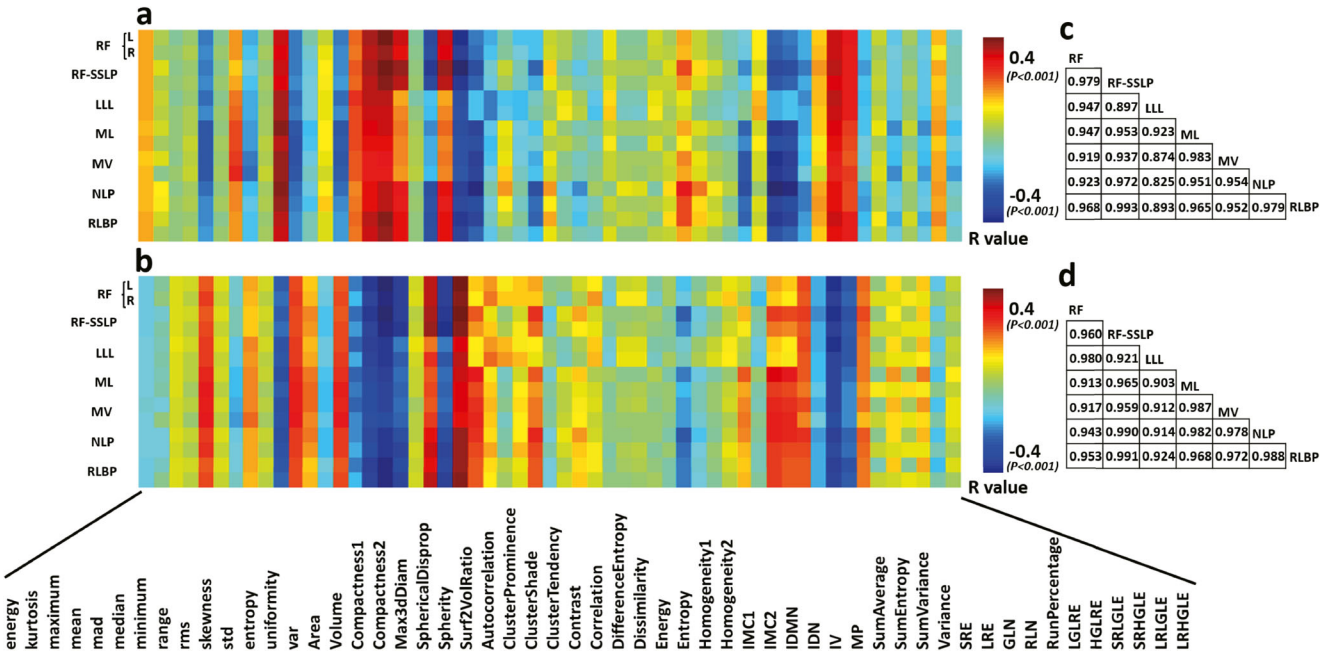

Fig. 4 The correlation map between the radiomic features and clinical information. The *R*-values of the correlation between radiomic features and MMSE (**a**), ADAS-cog13 (**b**) were calculated based on seven hippocampal segmentation methods; **c** and **d** were the correlation of the *R*-value in **a** and **b** between the seven segmentation methods. L means left and R means right of the hippocampus. MMSE, mini-mental state examination; ADAS-cog13, Alzheimer's disease assessment scale

with different imaging protocols, which further demonstrated the reliability and robustness of the results particularly on ADNI cohort with such heterogeneity.

Classification performance

Based on a machine learning–based classification of AD and NC, AUC > 0.88 (maximum = 0.89) and ACC > 83.97% (maximum = 85.67%) were observed for all involved segmentation methods of MV, RF, RF-SSLP, ML, LLL, NLP, and RLBP when considering both left and right HRFs (Table 2, and Fig. 5). It is noted that the classification model was produced without using any feature selection method due to the low-dimension feature vector (N = 110).

When sufficient-segmenting or under-segmenting the hippocampus by MV with a series of threshold values (from 0.1 to 0.9, step size = 0.1), the ACC values ranged from 78.56 to 86.23% (combined left and right HRFs) were observed (Supplementary materials S05, and Fig. 5). The maximum ACC achieved when the hippocampus was sufficiently segmented by a threshold of 0.2, which was also the best performance when using HRFs in AD and NC classification.

Regarding the classification of AD and MCI, the bilateral HRFs exhibited AUC > 0.73 (maximum = 0.76) and ACC > 67.79% (maximum = 70.36%) for all involved segmentation methods. Besides, the performance of bilateral HRFs under sufficient-segmentation or under-segmentation showed the ACC ranged from 62.99 to 70.50%, and the AUC ranged from 0.68 to 0.77 (Supplementary materials S06). The maximum ACC and AUC were obtained by the sufficient-segmentation with a MV threshold of 0.1, which further proved the sufficient segmentation of the hippocampus could contribute to a better AD classification.

Consistency results were obtained with the EDSD dataset

The abovementioned results were reproduced in an independent dataset EDSD with n = 571 shown in the supplementary materials. Interestingly, the results showed very high consistency between the results in the ADNI dataset ("Classification performance" section) and the EDSD dataset (supplementary Table S2, Table S3, and Fig. S1 in supplementary materials S01).

Test-retest analysis to identify the most reliable radiomic features

A total of 110 (55 × 2) HRFs were extracted for each individual, of which, 74 features obtained ICC > 0.7 in ADNI, and 67 features in EDSD (Table 3). More importantly, 64 identical features are presented in both cohorts (Table 4), indicating the strong degree of reproducibility for the reliability of radiomic features. The ICCs for radiomic features in ADNI and EDSD cohorts are exhibited in the supplementary Fig. S15 and Fig. S16 in supplementary materials S07, respectively.

Table 2 The ACC, SPE, SEN, and AUC of the classification AD and NC in the ADNI cohort when the different segmentation methods were used. *ACC*, accuracy; *SPE*, specificity; *SEN*, sensitivity; *AUC*, the area under the receiver operating characteristic; *AD*, Alzheimer's disease; *NC*, normal control

Segmentation method	Feature	ACC	SPE	SEN	AUC
RF	Left	85.44%	91.87%	71.73%	0.89
	Right	84.13%	91.71%	68.55%	0.85
	Left+right	85.67%	92.70%	70.67%	0.89
RF-SSLP	Left	85.67%	92.54%	71.02%	0.89
	Right	84.99%	92.70%	68.55%	0.86
	Left+right	85.55%	92.70%	70.32%	0.89
LLL	Left	83.97%	90.05%	71.02%	0.88
	Right	83.30%	91.04%	66.78%	0.85
	Left+right	84.99%	92.37%	69.26%	0.88
ML	Left	84.09%	91.04%	69.26%	0.87
	Right	83.52%	91.21%	67.14%	0.85
	Left+right	84.31%	92.70%	66.43%	0.88
MV	Left	84.76%	91.54%	70.32%	0.87
	Right	83.63%	90.71%	68.55%	0.85
	Left+right	84.54%	92.04%	68.55%	0.88
NLP	Left	84.20%	91.87%	67.84%	0.88
	Right	84.42%	92.04%	68.20%	0.85
	Left+right	84.88%	92.54%	68.55%	0.89
RLBP	Left	84.99%	91.38%	71.38%	0.89
	Right	84.09%	91.38%	68.55%	0.86
	Left+right	85.55%	93.03%	69.61%	0.89

Discussion

In the present study, we investigated how different segmentation methods affect hippocampal radiomic feature accuracy in Alzheimer's disease analysis. Specifically, plenty of complicated MAIS methods for hippocampus segmentation were proposed and chasing the segmentation accuracy. Given the segmentation accuracy which was only slightly improved between different segmentation methods, we validated the measurement/statistical/clinical consistency of HRFs calculated from 7 different hippocampal segmentation methods. We concluded that HRFs showed highly consistency across different hippocampal segmentation methods. Besides, a machine learning–based classification of AD vs. NC and AD vs. MCI adopting the different HRFs demonstrated that the naïve MV (threshold value = 0.1 or 0.2) which produced a more sufficient segmentation with a relatively low segmentation accuracy was the best method to extract hippocampal radiomic features and achieved the highest accuracy in AD classification analysis. It indicated that the naïve MV method might be sufficient when using radiomics in diagnosing AD, and chasing the hippocampus segmentation accuracy with a complicated strategy might be unnecessary.

In recent studies, radiomic features played an important role in diagnosing disease [7, 23, 24]. Numerous studies emphasized the importance of radiomics [24], and considered radiomics as a bridge between medical imaging and personalized medicine [25]. More importantly, radiomics also increased the precision of diagnosis, treatment, and prognosis of the tumor [6, 26, 27]. In recent years, hippocampal radiomic features were confirmed to serve as a neuroimaging biomarker for AD [8–11, 28]. Our results confirmed and extended previous neuroimaging findings [8–11, 28], which showed robust AD-related alterations of radiomic features. Briefly, the intensity features including kurtosis, mean, mad, median, entropy, and uniformity and shape features including size, area, compactness, and surf2vol showed significant differences between the AD and NC groups. Interestingly, the textural features, including LRE, GLN, RLN, and SRE, were significantly different between the AD and NC groups, consistent with previous research [10, 11]. More importantly, the radiomic features, which showed significant differences between the AD and NC groups, were almost repeated using different segmentation methods. It is crucial for the research community to elucidate reproducible and replicative biomarkers for various diseases [29, 30]. This study proved that the radiomic features were robust between different hippocampal segmentation methods.

In recent years, the computer-aided diagnosis of AD was with around 90% accuracy [10, 31–34], of which, two critical components were considered in AD analysis in terms of sMRI: (1) gray matter volume of voxel or ROI with deep learning or traditional machine learning method [32, 35]; (2) the features derived from the ROI [9–11], particularly from the hippocampus. Furthermore, the hippocampal radiomic features were also suggested as sensitive neuroimaging biomarkers for AD [8–11, 28], which could obtain about 85–90% accuracy in the classification AD and NC. In general, the performance of the classification model was influenced by the accuracy of the hippocampal segmentation [18]. Interestingly, the accuracy of the hippocampal segmentation methods was varied from 85 to 90%, and the accuracy of the classification of AD and NC was only varied from 84 to 86% in Table 2 (68 to 70% in classification AD and MCI in Table S10) based on different hippocampal segmentation methods. Besides, we did not find a significant correlation between the accuracy of the hippocampal segmentation methods and the accuracy of AD and NC classification. Thus, the radiomic features were influenced little by different hippocampal segmentation methods.

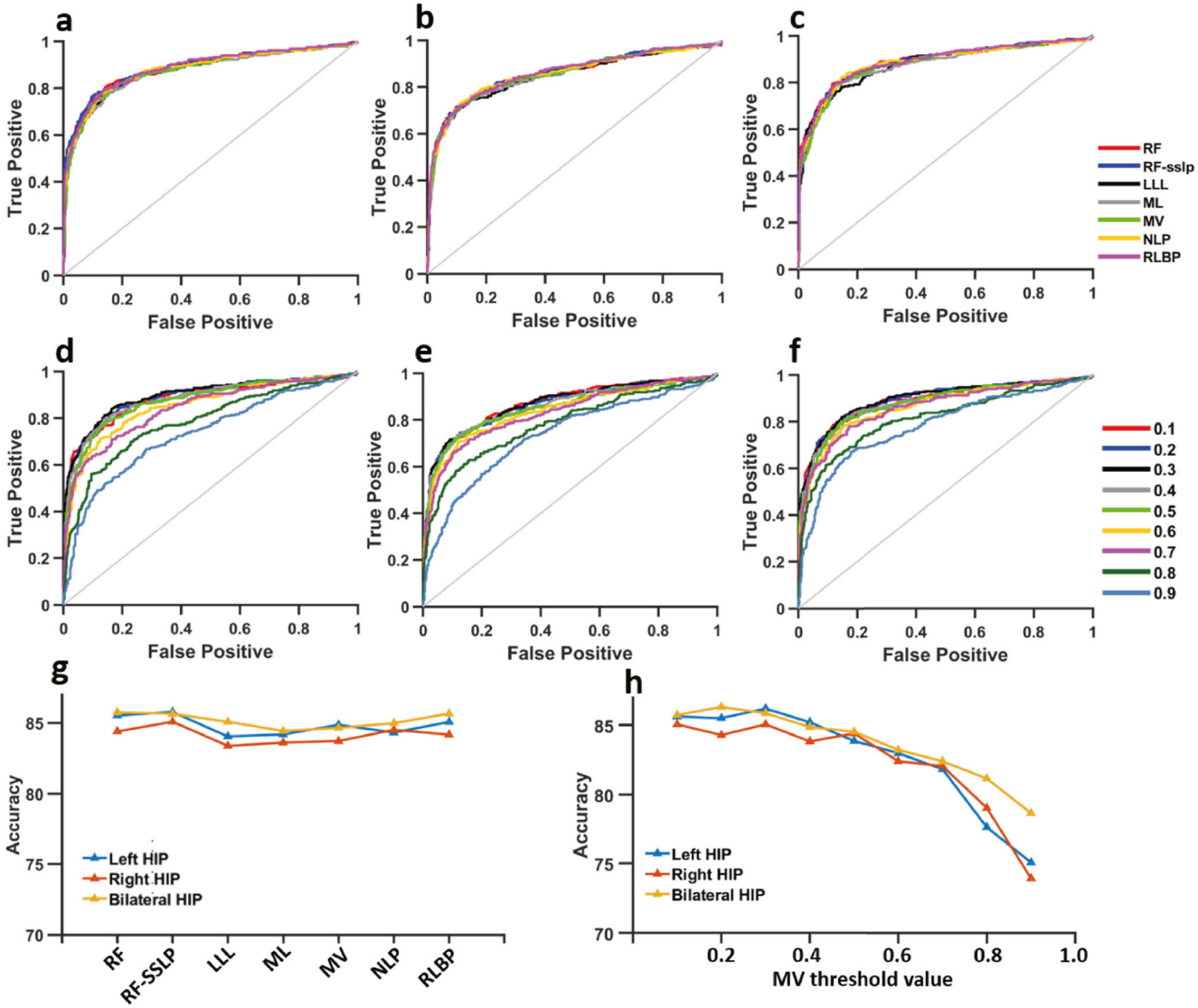

Fig. 5 The results of classification AD from NC obtained based on different segmentation methods. The ROC curve of the classification results was obtained from RF, RF-SSLP, LLL, ML, MV, NLP, and RLBP based on the bilateral hippocampus (**a**), left hippocampus (**b**), and right hippocampus (**c**). The ROC curve of the classification results was obtained from different threshold values (0.1–0.9 with step size = 0.1) based on the bilateral hippocampus (**d**), left hippocampus (**e**), and right hippocampus (**f**). The accuracy was obtained from different segmentation methods (**g**) and the different threshold values (**h**). RF, random forest; RF-SSLP, random forest semi-supervised label propagation; LLL, local label learning; ML, metric learning; MV, majority voting; NLP, nonlocal patch; RLBP, random local binary pattern

In our study, the classification model's accuracy between AD and NC was varied from 74 to 86% when the MV method was adopted in hippocampal segmentation with a series of threshold values. A more vast segmentation of the hippocampus was obtained with the threshold value = 0.1 (ACC = 86%). In contrast, a smaller region of the hippocampus was obtained with the threshold value = 0.9 (ACC = 74%). The MV method was a primitive MAIS method [18] with relatively low accuracy in hippocampal segmentation. However, the accuracy of classification AD and NC based

Table 3 The radiomic features with ICC > 0.7 in the ADNI and EDSD cohorts. *ICC*, intraclass correlation coefficient

	Left hippocampus			Right hippocampus		
	Intensity (14)	Shape (8)	Texture (33)	Intensity (14)	Shape (8)	Texture (33)
ADNI	14	8	15	14	8	15
EDSD	13	5	15	13	5	16

Table 4 Radiomic features with ICC > 0.7 in the ADNI and EDSD cohorts. Identical features are in bold. *ICC*, intraclass correlation coefficient

Intensity	**Median**	**Rms**	**Mean**	**Energy**
	Maximum	**Mad**	**Std**	**Var**
	Range	**Skewness**	**Minimum**	**Kurtosis**
	Entropy	Uniformity		
Shape	**Volume**	**Max3dDiam**	**Compactness1**	**Area**
	Surf2VolRatio	Spherity	SphericalDisprop	Compactness2
Texture	**GLN**	**RLN**	**SRE**	**LRE**
	RunPercentage	**LRLGLE**	**SRHGLE**	**HGLRE**
	LRGLE	**LGLRE**	**SRLGLE**	**SumVariance**
	Autocorrelation	**Variance**	SumAverage	Contrast

on MV in a typical setting of threshold value = 0.5 was almost equal with other hippocampal segmentation methods. Importantly, the highest ACC = 86% was obtained based on the threshold value of MV = 0.1 or 0.2. The result enlightened that hippocampal surrounding area covering the voxels with small gray matter volume at the edge of the hippocampus obtained by a more sufficient segmentation could also provide valuable information when extracting radiomic features for AD analysis. Moreover, the identified patterns were highly consistent within the AD vs. MCI classification, which demonstrated the reliability and reproducibility of the study findings (Supplementary materials S06).

For multi-atlas image segmentation (MAIS) methods, especially for the MV method, the larger the threshold, the smaller the segmented hippocampus (under segmentation), causing the loss of hippocampal information. In contrast, the lower threshold means that the hippocampus was segmented more completely (sufficient segmentation), and therefore preserves more useful information. As shown in Fig. 6a, the segmentation masks of the hippocampus obtained by the MV method increase with the threshold going smaller. However, the radiomics can only be calculated based on the entire segmentation region of the hippocampus, no matter sufficient segmentation or under segmentation. In this study, the best classification accuracy was obtained when the threshold value = 0.1 or 0.2, generating a sufficient segmentation of the hippocampus covering more surrounding voxels with small gray matter volume at the edge of the hippocampus. Therefore, we only concluded that the surrounding area of the hippocampus could contribute to AD analysis given the sufficient segmentation and its associated radiomics, but finding the specific radiomic feature that most correlated to the surrounding area was impracticable.

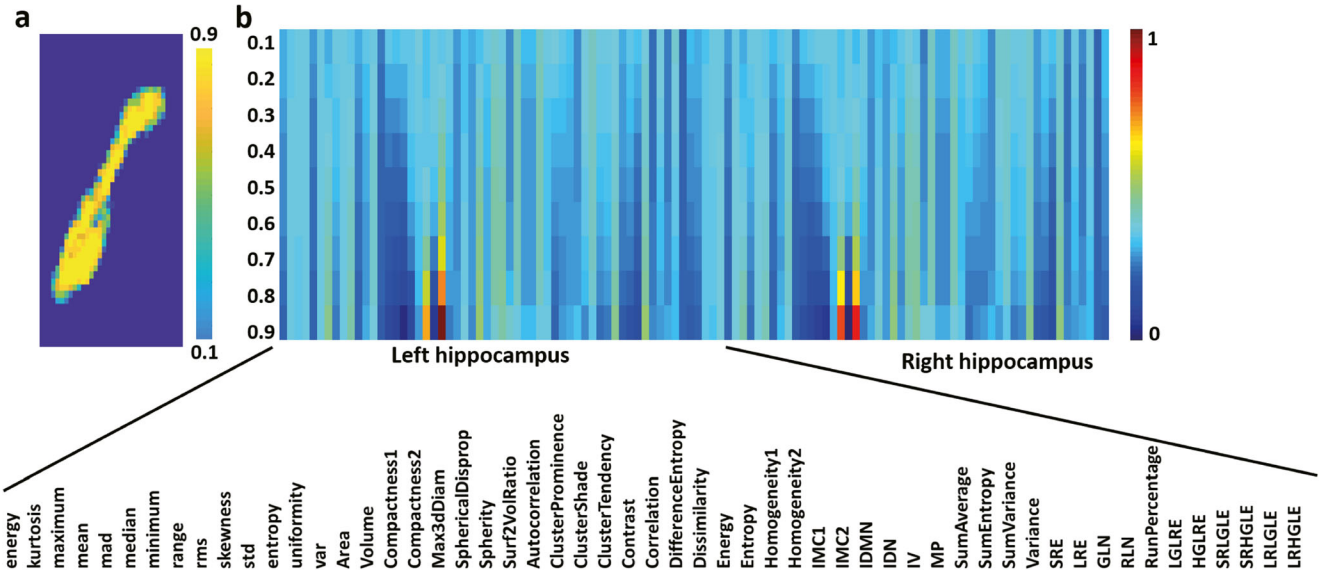

Fig. 6 The hippocampal segmentation result and radiomic features are obtained based on the MV. **a** An example of the hippocampal segmentation results obtained with the threshold value ranged from 0.1 to 0.9; **b** the mean value of the radiomic features (normalized) among all individuals based on the different threshold values

The comparison between multiple different segmentation algorithms, other than just one widely used segmentation technique, is essential in the present study. Although the typical MV method could generate different segmentations by a series of thresholds (Fig. 6a), but it is a naïve hippocampus segmentation method and achieves relatively low segmentation accuracy. The experiments solely on MV method could be insufficient to demonstrate the hypothesis in our study. Currently, plenty of advanced machine learning–based MAIS methods were proposed based on the MV [36], and were widely used to accurately segment the hippocampus. As such, it is necessary to compare them for exploring whether different segmentation methods would significantly affect HRFs when used in AD analysis. Besides, the segmentation accuracy was only slightly improved but at the expense of high computational cost; therefore, finding out optimal hippocampus segmentation strategy when extracting radiomics features was also of great importance in AD analysis.

There are some limitations to this study. First, the age and sex were not matched among the NC, MCI, and AD groups due to the data acquisition. Second, the study included only cases from the ADNI and EDSD databases with strictly controlled, and the result should be confirmed in the general clinical dataset with varied data acquisition parameters and image quality. Third, the patients, such as vascular co-morbidity, should be considered in the further study.

In conclusion, the results in our study demonstrated that HRFs exhibited high measurement/statistical/clinical consistency across different hippocampal segmentation methods, and the best performance in AD classification was obtained when HRFs were extracted by the naïve majority voting method with a more sufficient segmentation and relatively low hippocampus segmentation accuracy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09081-y.

Acknowledgements 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). The ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and generous contributions from 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 provided funds to support ADNI clinical sites in Canada. Private sector contributions were facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization was the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding This study has received funding by the National Natural Science Foundation of China (61802330, 61802331, 61801415).

Declarations

Guarantor The scientific guarantor of this publication is Qiang Zheng from Yantai University, the lead author of the study.

Conflict of interest The authors declare no competing interests.

Statistics and biometry One of the authors (Minhui Ouyang, the last author, Children's Hospital of Philadelphia) has significant statistical expertise.

Informed consent Informed written consent was obtained from all subjects (patients) in this study.

Ethics approval The study was approved by the institutional review boards of all the participating institutions.

Study subjects or cohorts overlap: It should be noted that 990 of the 1650 subjects in ADNI cohort have been previously reported (Zhao et al, Jin et al). These prior studies focused on whether the hippocampal radiomics feature (Zhao et al) or 3D attention network model (Jin et al) can distinguish AD from NC, whereas the present study aims to test how segmentation methods affect hippocampal radiomic feature accuracy in Alzheimer's disease analysis.

Zhao K, Ding Y, Han Y, et al Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis. Science Bulletin. 2020;65(13):1103-13.

Jin D, Wang P, Zalesky A, et al Grab-AD: Generalizability and reproducibility of altered brain activity and diagnostic classification in Alzheimer's Disease. Hum Brain Mapp. 2020;41(12):3379-91.

Methodology
- retrospective
- case-control study
- multicenter study

References

1. Handels RL, Wolfs CA, Aalten P, Joore MA, Verhey FR, Severens JL (2014) Diagnosing Alzheimer's disease: a systematic review of economic evaluations. Alzheimers Dement 10:225–237
2. Jack CR Jr, Albert MS, Knopman DS et al (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 7:257–262

3. Querfurth HW, LaFerla FM (2010) Alzheimer's disease. N Engl J Med 362:329–344
4. Petersen RC (2016) Mild cognitive impairment. Continuum (Minneap Minn) 22:404–418
5. Petersen RC, Roberts RO, Knopman DS et al (2009) Mild cognitive impairment: ten years later. Arch Neurol 66:1447–1455
6. Aerts HJ, Velazquez ER, Leijenaar RT et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006
7. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446
8. Chaddad A, Desrosiers C, Niazi T (2018) Deep radiomic analysis of MRI related to Alzheimer's disease. IEEE Access 6:58213–58221
9. Sorensen L, Igel C, Liv Hansen N et al (2016) Early detection of Alzheimer's disease using MRI hippocampal texture. Hum Brain Mapp 37:1148–1161
10. Zhao K, Ding Y, Han Y et al (2020) Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis. Sci Bull 65:1103–1113
11. Feng F, Wang P, Zhao K et al (2018) Radiomic features of hippocampal subregions in Alzheimer's disease and amnestic mild cognitive impairment. Front Aging Neurosci 10:290
12. Rohlfing T, Brandt R, Menzel R, Maurer CR Jr (2004) Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. Neuroimage 21:1428–1442
13. Coupe P, Manjon JV, Fonov V, Pruessner J, Robles M, Collins DL (2011) Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage 54:940–954
14. Zhu H, Tang Z, Cheng H, Wu Y, Fan Y (2019) Multi-atlas label fusion with random local binary pattern features: application to hippocampus segmentation. Sci Rep 9:16839
15. Zhu H, Cheng H, Yang X, Fan Y, Alzheimer's Disease Neuroimaging I (2017) Metric learning for multi-atlas based segmentation of hippocampus. Neuroinformatics 15:41–50
16. Hao Y, Wang T, Zhang X et al (2014) Local label learning (LLL) for subcortical structure segmentation: application to hippocampus segmentation. Hum Brain Mapp 35:2674–2697
17. Han X (2013) Learning-boosted label fusion for multi-atlas auto-segmentation. International Workshop on Machine Learning in Medical Imaging. Springer, pp 17–24
18. Zheng Q, Wu Y, Fan Y (2018) Integrating semi-supervised and supervised learning methods for label fusion in multi-atlas based image segmentation. Front Neuroinform 12:69
19. Mohs RC, Rosen WG, Davis KL (1983) The Alzheimer's disease assessment scale: an instrument for assessing treatment efficacy. Psychopharmacol Bull 19:448–450
20. Jack CR Jr, Bernstein MA, Fox NC et al (2008) The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 27:685–691
21. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420
22. Zhao K, Zheng Q, Che T et al (2020) Regional radiomics similarity networks (R2SNs) in the human brain: reproducibility, small-world properties and a biological basis. Netw Neurosci:1–30
23. Kumar V, Gu Y, Basu S et al (2012) Radiomics: the process and the challenges. Magn Reson Imaging 30:1234–1248
24. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. Radiology 278:563–577
25. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762
26. Coroller TP, Grossmann P, Hou Y et al (2015) CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 114:345–350
27. Huang X, Cheng Z, Huang Y et al (2018) CT-based radiomics signature to discriminate high-grade from low-grade colorectal adenocarcinoma. Acad Radiol 25:1285–1297
28. Feng Q, Ding Z (2020) MRI radiomics classification and prediction in Alzheimer's disease and mild cognitive impairment: a review. Curr Alzheimer Res 17:297–309
29. Freedman LP, Venugopalan G, Wisman R (2017) Reproducibility2020: progress and priorities. F1000Res 6:604
30. Velasco-Annis C, Akhondi-Asl A, Stamm A, Warfield SK (2018) Reproducibility of brain MRI segmentation algorithms: empirical comparison of Local MAP PSTAPLE, FreeSurfer, and FSL-FIRST. J Neuroimaging 28:162–172
31. Jin D, Wang P, Zalesky A et al (2020) Grab-AD: generalizability and reproducibility of altered brain activity and diagnostic classification in Alzheimer's disease. Hum Brain Mapp 41:3379–3391
32. Jin D, Zhou B, Han Y et al (2020) Generalizable, reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. Adv Sci (Weinh) 7:2000675
33. Li H, Habes M, Wolk DA et al (2019) A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimers Dement 15:1059–1070
34. Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C (2017) A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. Neuroimage 155:530–548
35. Beheshti I, Demirel H, Alzheimer's Disease Neuroimaging I (2016) Feature-ranking-based Alzheimer's disease classification from structural MRI. Magn Reson Imaging 34:252–263
36. Zheng Q, Wu Y, Fan Y (2018) Integrating semi-supervised and supervised learning methods for label fusion in multi-atlas based image segmentation. Front Neuroinf 12:69