European Radiology
https://doi.org/10.1007/s00330-023-09519-x

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

NEURO

A comprehensive characterization of hippocampal feature ensemble serves as individualized brain signature for Alzheimer's disease: deep learning analysis in 3238 participants worldwide

Yiyu Zhang[1] · **Hongming Li**[2] · **Qiang Zheng**[1]

Received: 10 August 2022 / Revised: 19 December 2022 / Accepted: 2 February 2023

Abstract

Objectives Hippocampal characterization is one of the most significant hallmarks of Alzheimer's disease (AD); rather, the single-level feature is insufficient. A comprehensive hippocampal characterization is pivotal for developing a well-performing biomarker for AD. To verify whether a comprehensive characterization of hippocampal features of gray matter volume, segmentation probability, and radiomics features could better distinguish AD from normal control (NC), and to investigate whether the classification decision score could serve as a robust and individualized brain signature.

Methods A total of 3238 participants' structural MRI from four independent databases were employed to conduct a 3D residual attention network (3DRA-Net) to classify NC, mild cognitive impairment (MCI), and AD. The generalization was validated under inter-database cross-validation. The neurobiological basis of the classification decision score as a neuroimaging biomarker was systematically investigated by association with clinical profiles, as well as longitudinal trajectory analysis to reveal AD progression. All image analyses were performed only upon the single modality of T1-weighted MRI.

Results Our study exhibited an outstanding performance (ACC = 91.6%, AUC = 0.95) of the comprehensive characterization of hippocampal features in distinguishing AD ($n = 282$) from NC ($n = 603$) in Alzheimer's Disease Neuroimaging Initiative cohort, and ACC = 89.2% and AUC = 0.93 under external validation. More importantly, the constructed score was significantly correlated with clinical profiles ($p < 0.05$), and dynamically altered over the AD longitudinal progression, provided compelling evidence of a solid neurobiological basis.

Conclusions This systemic study highlights the potential of the comprehensive characterization of hippocampal features to provide an individualized, generalizable, and biologically plausible neuroimaging biomarker for early detection of AD.

Key Points

- *The comprehensive characterization of hippocampal features exhibited ACC = 91.6% (AUC = 0.95) in classifying AD from NC under intra-database cross-validation, and ACC = 89.2% (AUC = 0.93) in external validation.*
- *The constructed classification score was significantly associated with clinical profiles, and dynamically altered over the AD longitudinal progression, which highlighted its potential of being an individualized, generalizable, and biologically plausible neuroimaging biomarker for early detection of AD.*

Keywords Alzheimer's disease · Hippocampus · Magnetic resonance imaging · Deep learning · Residual attention network

Abbreviations

3DRA-Net	3D residual attention network
ACC	Accuracy
AD	Alzheimer's disease
ADAS	Alzheimer's Disease Assessment Scale
ADNI	Alzheimer's Disease Neuroimaging Initiative database
AIBL	Australian Imaging Biomarkers and Lifestyle Study of Aging
ANTs	The Advanced Normalization Tools
APOE	Apolipoprotein E

✉ Qiang Zheng
zhengqiang@ytu.edu.cn

[1] School of Computer and Control Engineering, Yantai University, No. 30, Qingquan Road, Laishan District, Yantai City 264005, Shandong Province, China

[2] Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Published online: 09 March 2023

AUC	Area under the ROC curve
Aβ	β-Amyloid
CDRSB	Clinical dementia rating sum of boxes
CSF	Cerebrospinal fluid
EDSD	European DTI Study on Dementia database
FAQ	Functional assessment questionnaire
FDG	Fluorodeoxyglucose
MCI	Mild cognitive impairment
MMSE	Mini-Mental State Examination
MNI	Montreal Neurological Institute
MRI	Magnetic resonance imaging
NC	Normal control
NFTs	Neurofibrillary tangles
OASIS	Open Access Series of Imaging Studies
PET	Positron emission tomography
PHS	Polygenic hazard score
pMCI	Progressive MCI
P-Tau	Tau phosphorylated at threonine 181
Ravlt	Rey auditory verbal learning test
Res block	Residual block
RF-SSLP	Random forest-semi-supervised label propagation
ROC	Receiver operating characteristic
SEN	Sensitivity
sMCI	Stable MCI
SPE	Specificity
VBM	Voxel-based morphometric method

Introduction

Alzheimer's disease (AD), the most prevalent cause of dementia in the elderly, is an irreversible neurodegenerative disorder characterized by progressive cognitive impairment and functional deterioration [1]. Mild cognitive impairment (MCI) is usually considered the prodromal stage for AD [2], and the intervention for preclinical AD and MCI is the most efficient way to delay the decline of cognition and progression of AD pathology [3]. Thus, establishing clinically available biomarkers that can accurately identify individuals with preclinical or early stages of AD is particularly important for clinical intervention and precise medicine. However, the existing biomarkers are not sufficiently convenient and generalizable.

Hippocampal characterization can be served as one of the most significant hallmarks of AD [4]. Previous studies have achieved high accuracy for distinguishing AD from normal controls (NCs) based on the shape or texture features of the hippocampus [5, 6]. It is well accepted that the hippocampus is a complex system for supporting the integration of spatial information and memory encoding [7]. However, the single-level features such as gray matter volume is insufficient to characterize hippocampus as a generalizable biomarker. That also explains why many studies turn to characterizing the whole brain by an interpretable deep learning [8–10]. It is worth noting that the deep learning model particularly for the training procedure established on the whole brain suffers the computational cost. Therefore, a well-performing neuroimaging biomarker for AD established on deep learning and a comprehensive characterization of hippocampal feature ensemble is expected.

Abnormal patterns of the hippocampus can be unveiled from different perspectives. For instance, the gray matter volume which directly characterizes the hippocampal atrophy severity is one of the most reliable biomarkers for AD [11]. The probability matrix, which is a probabilistic segmentation map produced by the multi-atlas hippocampus segmentation method with each value being the probability of corresponding voxel belonging to the hippocampus [12], has also been validated as a promising predictor of AD progression [13]. The hippocampal radiomics features have also been proved to be encouraging biomarkers for AD, including intensity, shape, and textural features [14]. Thus, we hypothesize the biomarker devised by a comprehensive characterization of hippocampal feature ensemble of gray matter volume, segmentation probability matrix, and radiomics might obtain an ideal performance than that only with single-level features.

The first aim of this study is to verify whether the comprehensive characterization of hippocampal feature ensemble could better serve to distinguish AD from NC. The second aim is to investigate whether an individual classification score could serve as a robust and biological neuroimaging biomarker. For this purpose, we first proposed a multi-feature ensemble classification model with 3D residual attention network (3DRA-Net) based on Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL), the European DTI Study on Dementia (EDSD), and the Open Access Series of Imaging Studies (OASIS) cohorts ($n = 3238$). Then, we investigated whether the neuroimaging biomarker derived from the classification decision score has a solid neurobiological basis by relating it with clinical profiles (e.g., mini-mental state examination (MMSE), apolipoprotein E (APOE) genotype, polygenic hazard score (PHS), fluorodeoxyglucose (FDG), cerebrospinal fluid (CSF) Aβ, CSF Tau, and other clinical measures). At last, the longitudinal trajectory study of distinct measures was further performed to evaluate whether this biomarker could dynamically change to track the disease progression. The schematic illustration of the deep learning–based AD analysis is summarized in Fig. 1. Preliminary results of this study have been reported in a conference paper [15].

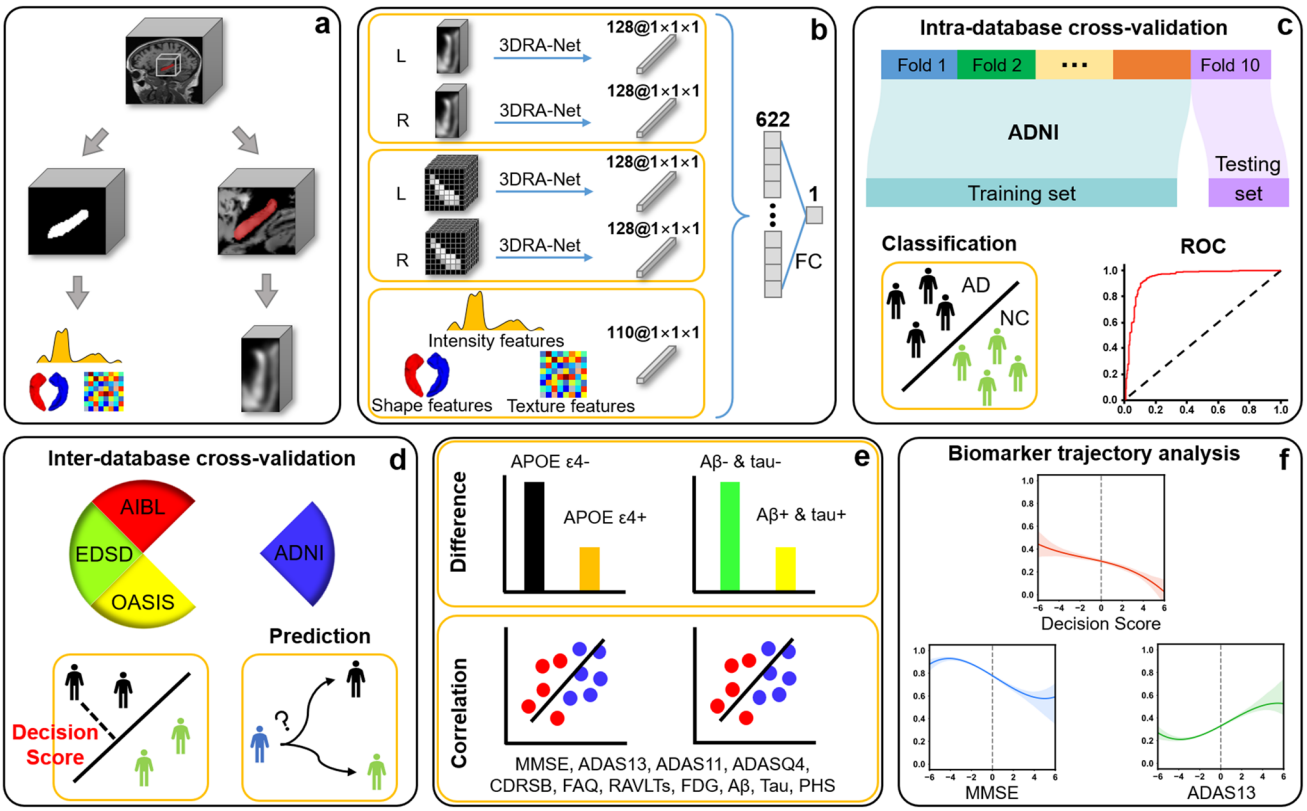

Fig. 1 Schematic illustration of the deep learning–based AD analysis. **a** Pipeline of the data preprocessing. **b** Framework of the multi-view hippocampal features ensemble. **c** Classification analysis of AD and NC under intra-database cross-validation in the ADNI cohort. **d** Inter-database cross-validation in diagnosing AD based on four independent cohorts, and prediction of MCI progressing to AD within 3 years in the ADNI cohort. **e** Statistical analysis of group differences and correlations between the decision score and clinical profiles. **f** Longitudinal trajectory analysis of the decision score, MMSE, and ADAS13 scores during the AD progression

Materials and methods

Data acquisition

A total of 3238 participants' structural MRI images including baseline T1-weighted scans were obtained from four independent cohorts: ADNI (http://adni.loni.usc.edu), AIBL (http://aibl.csiro.au), the EDSD (http://neugrid4you.eu), and the OASIS (http://oasis-brains.org) databases (Table 1 and Supplementary Materials S01). In the ADNI cohort (1649 participants), a total of 1267 participants (3006 scans) with a mean follow-up period of 4.29 ± 3.46 years were also included. It should be noted that the ADNI cohort was served as the primary discovery cohort due to its detailed clinical information.

Data preprocessing and feature extraction

The gray matter volume of the whole brain for each T1 MRI scan was computed via the CAT12 toolkit (http://dbm.neuro.

Table 1 Demographic summary about the subjects for all the independent cohorts

Cohort	Group	Age (years)	Sex (M/F)	MMSE
ADNI ($N=1649$)	NC (603)	73.46 ± 6.16	277/326	29.08 ± 1.10
	MCI (764)	72.98 ± 7.68	447/317	27.56 ± 1.81
	AD (282)	74.91 ± 7.69	151/131	23.18 ± 2.13
	p value	< 0.001	< 0.001	< 0.001
AIBL ($N=412$)	NC (334)	73.28 ± 5.94	140/194	28.67 ± 1.27
	AD (78)	74.33 ± 7.70	33/45	20.59 ± 5.28
	p value	0.283	0.950	< 0.001
EDSD ($N=388$)	NC (230)	68.76 ± 6.14	108/122	28.58 ± 2.97
	AD (158)	75.54 ± 8.10	66/92	20.89 ± 5.12
	p value	< 0.001	0.072	< 0.001
OASIS ($N=789$)	NC (599)	67.14 ± 8.72	244/355	29.06 ± 1.22
	AD (190)	74.99 ± 7.69	96/94	24.47 ± 4.13
	p value	< 0.001	0.018	< 0.001

uni-jena.de/cat/) [16]. To reduce the computational complexity, we defined a bounding box (60×48×60) to entirely cover the hippocampus (Fig. 1a). It should be noted that the size of bounding box was further cut into 62×26×38 to remove the influence of redundant voxels, and then followed by smooth processing.

The segmentation probability matrix (60×48×60) of the hippocampus was obtained based on the Random Forest-Semi-supervised Label Propagation algorithm [12] after performing N4 correction and linearly aligning all T1 MRI scans to the Montreal Neurological Institute space ($1\times1\times1$ mm^3) with the Advanced Normalization Tools (ANTs) (https://github.com/ANTsX/ANTs) [17], selecting 20 most similar atlases and non-linearly aligning them to the target image.

The radiomics features (intensity features ($n=14$), shape features ($n=8$), and textural features ($n=33$)) for each side hippocampus were computed (https://github.com/YongLiulab) [14] (Supplementary Materials S02).

Individual score from multi-feature ensemble classification model

In this study, a 3DRA-Net was developed to learn the most representative features for gray matter volume and probability matrix of the bilateral hippocampus (Figs. 1b and 2a). Then, an individual score (referred as decision score) was generated by integrating outputs of the 3DRA-Net (128×4) and hippocampal radiomics features (55×2) using fully connected layers for the individualized prediction (Fig. 1b).

In the 3DRA-Net, the attention mechanism consisting of a trunk branch and a soft mask branch was to guide feature learning in an end-to-end training fashion (Fig. 2b). Moreover, a novel residual block (Res block) was proposed and integrated into down-sampling pathway to boost the capacity of representation and overcome the gradient disappearance (Fig. 2c). The global average pooling was adopted to compress all voxels of single-channel feature map into one value to avoid over-fitting.

Due to the imbalance of AD and NC in present study, a threshold strategy was combined with the loss function of binary cross-entropy to address such issue, which redefine the typical 0.5 threshold to a proportion of AD/(AD+NC) in the training set. Besides, the proposed 3DRA-Net was implemented using the platform of Pytorch (version=1.7.1), which was initialized with He's initialization [18], and optimized using the optimizer Adam [19] with initial learning rate of 3×10^{-4}, weight decay rate of 1×10^{-4}, and mini-batch size of 8. The dropout rate before the fully connected layers was 0.3, and the training time was 70 epochs.

Classification analysis, validation, and generalizability

We first constructed the classification model for classifying AD ($n=282$) and NCs ($n=603$) based on different hippocampal features in the ADNI cohort under tenfold cross-validation (Fig. 1c). In short, the participants were

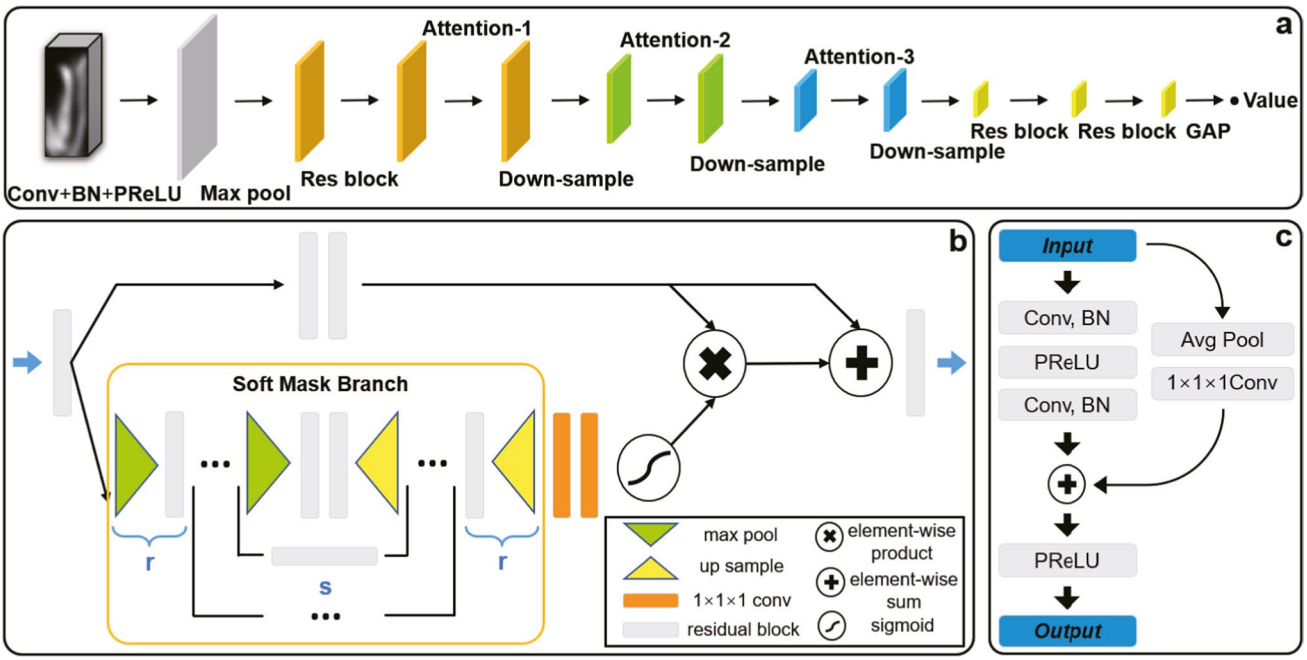

Fig. 2 Architecture of the 3D convolutional neural network. **a** 3D residual attention network. **b** Residual attention module. The parameters r and s in the soft mask branch of the "attention-1," "attention-2," and "attention-3" modules in the 3DRA-Net are 3/2, 2/1, and 1/0, respectively. **c** Residual block

split into training sets (80% of the data), a validating set (10% of the data), and a testing set (10% of the data). In addition, to further assess the robustness of the classification model, the inter-database cross-validation with leave-center-out strategy was considered based on four independent cohorts, including ADNI (603 NCs, and 282 AD), AIBL (334 NCs, and 78 AD), EDSD (230 NCs, and 158 AD), and OASIS (599 NCs, and 190 AD) (Fig. 1d). Briefly, one independent cohort was being the testing set in turn, while the remaining three cohorts were split into training sets (90% of the data), and a validating set (10% of the data) (Supplementary Materials S03).

Furthermore, to discriminate the progressive MCI patients (pMCI) from stable MCI patients (sMCI), the above AD/NC classifier trained on ADNI (not training another classifier with MCI subjects) was directly employed to predict whether the MCI progress to AD within 3 years in the ADNI cohort (150 pMCI, and 252 sMCI) (Fig. 1d and Supplementary Materials S04). Besides, we also compared the AUC of the decision score and other clinical measures in predicting MCI convert to AD (58 pMCI, and 138 sMCI) (Supplementary Materials S05).

Statistical analysis

To assess the decision score in different groups, a two-sample two-sided *t*-test was performed among the NC, sMCI, pMCI, and AD groups. As the APOE gene is a significant genetic risk factor for AD [20, 21], we further explored the difference of decision scores between subjects with APOE ε4+ and APOE ε4− in the NC, MCI, and AD groups, respectively. Besides, the combined presence of Aβ plaques and Tau neurofibrillary tangles (NFTs) is a unique hallmark of AD [22]. Thus, the group difference analysis of the decision scores was also performed among three MCI subgroups (Aβ+&Tau+, Aβ+&Tau−/(Aβ−&Tau+), and Aβ−&Tau−) (Fig. 1e). Here, Aβ+ was defined when Aβ < 1098 pg/mL, and Tau+ was defined when Tau > 242 pg/mL [23, 24].

We also explored the biological basis of the decision score by relating it to the clinical measures, such as cognitive ability (MMSE, ADAS13, ADAS11, ADASQ4, Ravlt-immediate, Ravlt-learning, CDRSB, and FAQ), CSF biomarker (CSF Aβ, and CSF Tau), metabolism (FDG), and genetic risk score (PHS). The correlation analysis was performed in the MCI and AD groups after controlling for the effects of age, gender, and clinical group (Fig. 1e).

Progression trajectory analysis of the individual score, MMSE, and ADAS13 changes

Identification of dynamic changes for biomarkers during the AD progression is crucial for defining the disease stage and monitoring the efficacy of potential treatments [25]. To test the overall consistent patterns of longitudinal progression of the proposed biomarker and cognitive ability, the longitudinal trajectory analysis of the decision score, MMSE, and ADAS13 scores was performed in the NC, sMCI, pMCI, and AD groups in the ADNI cohort (Fig. 1f).

In particular, for the NC, sMCI, and AD subjects, the status remained stable until the last visit, and the baseline was therefore set as the origin of progressing time. However, the origin of pMCI subjects was set by the AD onset time point; i.e., the "0" point was defined as the time when the patient converted to AD. Then, the longitudinal trajectory was derived in each group using both linear and second-order non-linear regression models according to the decision score, MMSE, and ADAS13 scores of all subjects at each time point. The values of the decision score, MMSE, and ADAS13 scores were normalized by a max–min standardized method across all subjects.

Contribution of the decision score to diagnostic and predictive potential

We combined the clinical biomarkers and decision score to validate whether it could improve the overall diagnostic and predictive performance of AD. In particular, to preserve the advantage of the single modality of structural MRI (sMRI), a total of 11 indicators were considered, including decision score, age, gender, and 8 most easily accessible cognitive profiles (MMSE, ADAS13, ADAS11, ADASQ4, Ravlt-immediate, Ravlt-learning, CDRSB, and FAQ). The decision score for all subjects in contribution analysis was obtained by the 3DRA-Net under inter-database cross-validation.

The contribution analysis was conducted in two strategies according to study [26]: strategy (1)—the grid search algorithm was capitalized to excavate the best feature combination via the accuracy of testing set (referred as best-model-fit features). Strategy (2)—all the involved indicators were combined to evaluate the performance. For comparison, the decision score was removed in both strategies to assess the contribution of such score to the overall diagnostic predictive potential.

Under each strategy, we adopted a linear support vector machine (SVM) to implement a predictive experiment of whether MCI converts to AD within 3 years in the ADNI cohort, and a diagnostic experiment of AD and NC classification under inter-database cross-validation. Since the databases of AIBL, EDSD, and OASIS did not provide abundant indicators as ADNI, the decision score was only combined with age, gender, and MMSE in the classification of AD and NC under inter-database cross-validation experiment.

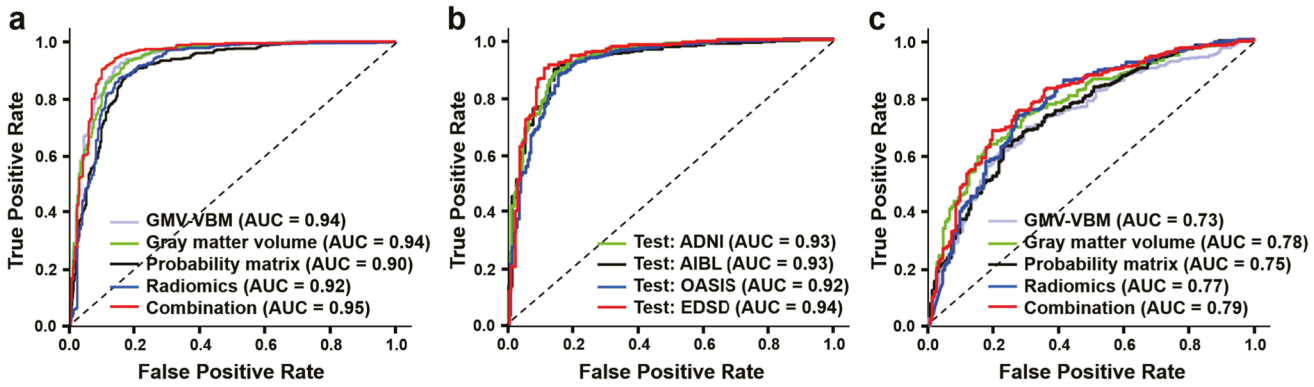

Fig. 3 ROC curves for the classification performance. **a** Classification of AD and NC with different hippocampal features under intra-database cross-validation in the ADNI cohort. **b** Inter-database cross-validation of the 3DRA-Net with multi-view ensemble hippocampal features based on four independent cohorts. **c** Prediction of MCI converting to AD within 3 years with different hippocampal features in the ADNI cohort

Results

Demographic characteristics and neuropsychological assessment

In total, 3238 subjects from four independent cohorts of ADNI ($n=1649$), AIBL ($n=412$), EDSD ($n=388$), and OASIS ($n=789$) were employed (Table 1 and Supplementary Materials S01). The MMSE score was significantly different among the NC, MCI, and AD groups in ADNI ($p<0.001$, ANOVA test). Similarly, a significant difference in the MMSE score was also observed between the NC and AD groups in AIBL, EDSD, and OASIS ($p<0.001$, t-test). The detailed clinical information is shown in Table 1.

Diagnostic performance

Regarding the classification of AD and NC, the multi-feature ensemble classifier yielded an ACC = 91.6% (SPE = 95.4%, SEN = 83.2%, AUC = 0.95) in the ADNI cohort, which is higher than that only with the single-level hippocampal features (Fig. 3a and Table 2). More importantly, we achieved a mean classification ACC = 89.2% (SPE = 92.3%, SEN = 80.5%, AUC = 0.93) with the comprehensive characterization of hippocampal feature ensemble under inter-database cross-validation based on four cohorts (Fig. 3b and Table 3).

Regarding the discrimination of pMCI from sMCI with baseline data in the ADNI cohort using the above AD/NC classifier (not training another classifier with MCI subjects) (150 pMCI, and 252 sMCI), the results showed the comprehensive characterization of hippocampal feature ensemble achieved an ACC = 74.8% (SPE = 80.6%, SEN = 65.1%, AUC = 0.79) (Fig. 3c and Supplementary Materials S04).

Besides, the comparison of the clinical measures and decision score in classifying pMCI and sMCI was also performed after excluding those MCI participants without complete clinical measures (mainly lack of the biochemical indicators) in the ADNI cohort (58 pMCI, and 138 sMCI). The experimental results demonstrated that the decision score (AUC = 0.79) outperformed other single clinical indicators, particularly for the CSF Aβ (AUC = 0.76), CSF Tau (AUC = 0.29), and CSF P-Tau (AUC = 0.28) (Supplementary Materials S05).

Table 2 Comparison of different hippocampal features in classifying AD and NC under intra-database cross-validation in the ADNI cohort. *VBM* voxel-based morphometric method

Feature	ACC (%)	SEN (%)	SPE (%)	AUC
Gray matter volume-VBM	89.6	82.9	92.6	0.94
Gray matter volume	89.8	79.3	94.6	0.94
Probability matrix	86.9	76.1	91.8	0.90
Radiomics	88.2	72.9	95.2	0.92
Combination	91.6	83.2	95.4	0.95

Comparison with other methods

The classification result of the present method was compared to other existing methods, which also employed deep learning on the baseline sMRI of whole brain [8–10] or hippocampus [27, 28] from ADNI. The results showed the comprehensive characterization of hippocampal feature ensemble achieved competitive performance (ACC = 91.6%, AUC = 0.95) compared with others (ACC ranging between 79.9 and 92.1%, AUC ranging between 0.86 and 0.94) (Supplementary Materials S06). Although the study [10] achieved the best ACC of 92.1%, the performance under inter-database cross-validation between ADNI and in-house was 86.1–87.0%, which was lower than our method (ACC = 89.2%). Of note, the performance of different methods [8–10, 27, 28] under comparison

Table 3 Classification performance of the 3DRA-Net with comprehensive characterization of hippocampal feature ensemble under inter-database cross-validation

Training set	Testing set	ACC (%)	SEN (%)	SPE (%)	AUC
AIBL + EDSD + OASIS	ADNI	88.4	81.9	91.4	0.93
ADNI + EDSD + OASIS	AIBL	89.8	82.1	91.6	0.93
ADNI + AIBL + EDSD	OASIS	89.4	74.4	94.2	0.92
ADNI + AIBL + OASIS	EDSD	88.7	83.5	92.2	0.94
Average		89.2	80.5	92.3	0.93

in Supplementary Materials S06 was directly cited from the corresponding studies due to the impractical identification of the same data from ADNI.

Furthermore, three deep learning models of FCN [9], 3DAN [10], and 3DResNet [29] were adopted to compare with the proposed 3DRA-Net under the intra- and inter-database cross-validations (Table 4). It is worth noting that the voxel-based morphometric (VBM) method was also involved as a baseline comparison method, which was implemented on the gray matter volume after a *t*-statistic-based feature selection (top 10,000 voxels were selected, Supplementary Materials S07) [30, 31].

The result showed that our method exhibited outstanding discrimination of AD (ACC = 89.2%, AUC = 0.93) in the external validations, which significantly outperformed the VBM (ACC = 83.1%, AUC = 0.91), FCN (ACC = 88.5%, AUC = 0.92), 3DAN (ACC = 87.2%, AUC = 0.92), and 3DResNet (ACC = 87.9%, AUC = 0.93) models (Table 4), and presented a better generalization. The results of the intra-database cross-validation in the ADNI cohort are listed in Supplementary Materials S06. Besides, all deep learning models under comparison achieved a high ACC (> 87% for all) with the external validation, further demonstrating the superiority of the comprehensive characterization of hippocampal feature ensemble in classifying AD and NC.

Associations between the decision score and clinical measures

In ADNI cohort, the significant difference was observed among NC, sMCI, pMCI, and AD groups ($p < 0.001$) (Fig. 4a). Besides, a significant difference between the individuals with or without APOE ε4+ in the MCI group was also obtained ($p < 0.001$) (Fig. 4b). Likewise, we also found a significant difference in the decision scores among three MCI subgroups (Aβ+&Tau+, Aβ+&Tau−/(Aβ−&Tau+), and Aβ−&Tau−) ($p < 0.001$) (Fig. 4c). The quantitative results are in Supplementary Materials S08.

Furthermore, Pearson's correlation exhibited that the decision score was significantly correlated with clinical measures (including MMSE, ADAS13, ADAS11, ADASQ4, Ravlt-immediate, Ravlt-learning, CDRSB, FAQ, FDG, CSF Aβ, and PHS in Fig. 5(a–k)) with all p values < 0.001 except CSF Tau. The correlation analysis conducted solely in the MCI or AD group are in Supplementary Materials S09.

Table 4 Comparison of the 3DRA-Net with other models in classifying AD and NC with comprehensive characterization of hippocampal feature ensemble under inter-database cross-validation

Method	ACC (%)	SEN (%)	SPE (%)	AUC
VBM	83.1	78.4	84.6	0.91
FCN	88.5	74.7	94.4	0.92
3DAN	87.2	75.4	92.4	0.92
ResNet	87.9	71.8	94.2	0.93
3DRA-Net	89.2	80.5	92.3	0.93

Group progress trajectories of the decision score, MMSE, and ADAS13

The MMSE and ADAS13 scores have been widely used to monitor the AD progression [25, 32, 33]. Herein, we also compared the longitudinal progression of the decision score to MMSE/ADAS13.

As shown in Fig. 6, the decision score remained relatively stable in the NC or sMCI group. However, the progressing trajectory of decision score gradually declined over time in the pMCI group (the lower the value, the higher the risk), which showed a high consistency with that of the MMSE (downward trend) and ADAS13 (upward trend) scores. Moreover, the second-order non-linear regression experiments also exhibited the same tendency between the decision score and MMSE/ADAS13 (Fig. S5 in Supplementary Materials S10), highlighting that the proposed neuroimaging marker was sensitive in delineating AD neurodegeneration. At last, Pearson's correlation also suggested the decision score was significantly correlated with MMSE/ADAS13 in the longitudinal trajectory for each clinical group ($p < 0.01$, Supplementary Materials S10), which statistically indicated the high trajectory consistency of longitudinal progression of AD between the decision score and cognitive ability.

Contribution of the individual score to the clinical biomarkers

In the contribution analysis (Table 5), the strategy (1) suggested that if the decision score was removed from

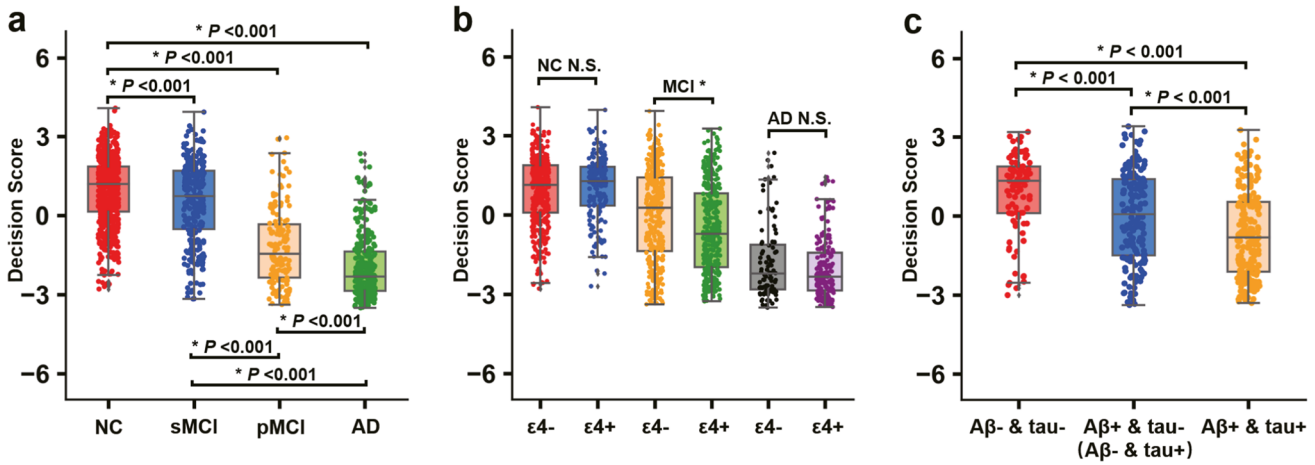

Fig. 4 Statistical analysis results of group differences in the ADNI cohort. **a** The decision score of subjects in the NC, sMCI, pMCI, and AD groups. **b** The decision score of subjects with or without APOE ε4+ in the NC, MCI, and AD groups. **c** The decision score of subjects with Aβ−&Tau−, Aβ+&Tau−/(Aβ−&Tau+), or Aβ+&Tau+in the MCI group. N.S., not significant

the best-model-fit features, the diagnostic and predictive accuracy would decrease by around 3%, and the sensitivity decreased by around 9–12%. Moreover, the strategy (2) suggested that the diagnostic and predictive accuracy would decrease by around 1–4%, and the sensitivity decreased by around 4–14% when the decision score was removed from all the involved indicators. In summary, these outcomes demonstrated that the decision score significantly contributed to the overall diagnostic predictive potential of AD.

Discussion

This study demonstrated that the comprehensive characterization of hippocampal feature ensemble could serve as robust and biological neuroimaging biomarkers for AD using intra- and inter-database cross-validations with deep learning techniques across four cohorts (ADNI, AIBL, EDSD, and OASIS ($n = 3238$)). Further, the association analyses between the constructed decision score and clinical profiles (e.g., APOE, CSF Aβ, and cognitive ability), as well as the longitudinal trajectory study of different measures during the AD progression, provided compelling evidence of a solid neurobiological basis. These findings highlight that our approach holds the potential to substantially drive early detection, progression monitoring, and therapeutic intervention for AD.

The establishment of valid biomarkers provides a strong endorsement in facilitating individual-specific therapies for AD. The present study integrated hippocampal gray matter volume, probability matrix, and radiomics features into an individual biomarker, which significantly improved the diagnostic reliability of AD. In our study, the gray matter volume achieved the best diagnosis among the classification models based on the single-level features. It should be noted that this result only indicated that the performance of the hippocampal volume was better than radiomics features and probability matrix in late-stage AD. Convergence studies also highlighted that high-order features, e.g., radiomics features, might obtain ideal performance in investigating the early stage of AD than gray matter volume [14, 34–37]. It is why we comprehensively combined the multi-view hippocampal features for the classification of AD and NC.

Generalization is the cornerstone of biomarkers being applied in distinct clinical environments [38]. The internal cross-validation based on a single cohort is limited by smaller sample sizes, causing the low robustness to another independent dataset [10, 39]. As such, externally validating models is central to ensure the generalizability in translational neuroimaging. Importantly, our approach achieved a satisfactory accuracy of 89.2% within external cross-validation, although the heterogeneity among the cohorts was significant regarding their MRI protocols, inclusion criteria, and clinical study methods, further highlighting its strong generalizability in dealing with independent datasets.

Systematic analyses for the association between the proposed marker and clinical profiles grounded our computational predictions in the biological evidence. The individual biomarker derived from the deep learning framework was significantly different between the sMCI and pMCI groups, which suggests this biomarker is very sensitive in preclinical AD. The APOE ε4 allele is the most vital genetic factor causing AD risk [20, 21]. As expected, the decision score was significantly different between APOE ε4+ and APOE ε4− in the MCI subjects. In a word, this biomarker has a genetic basis. Aβ plaques and Tau NFTs are pathological hallmarks of AD [40–42], and the significant difference in

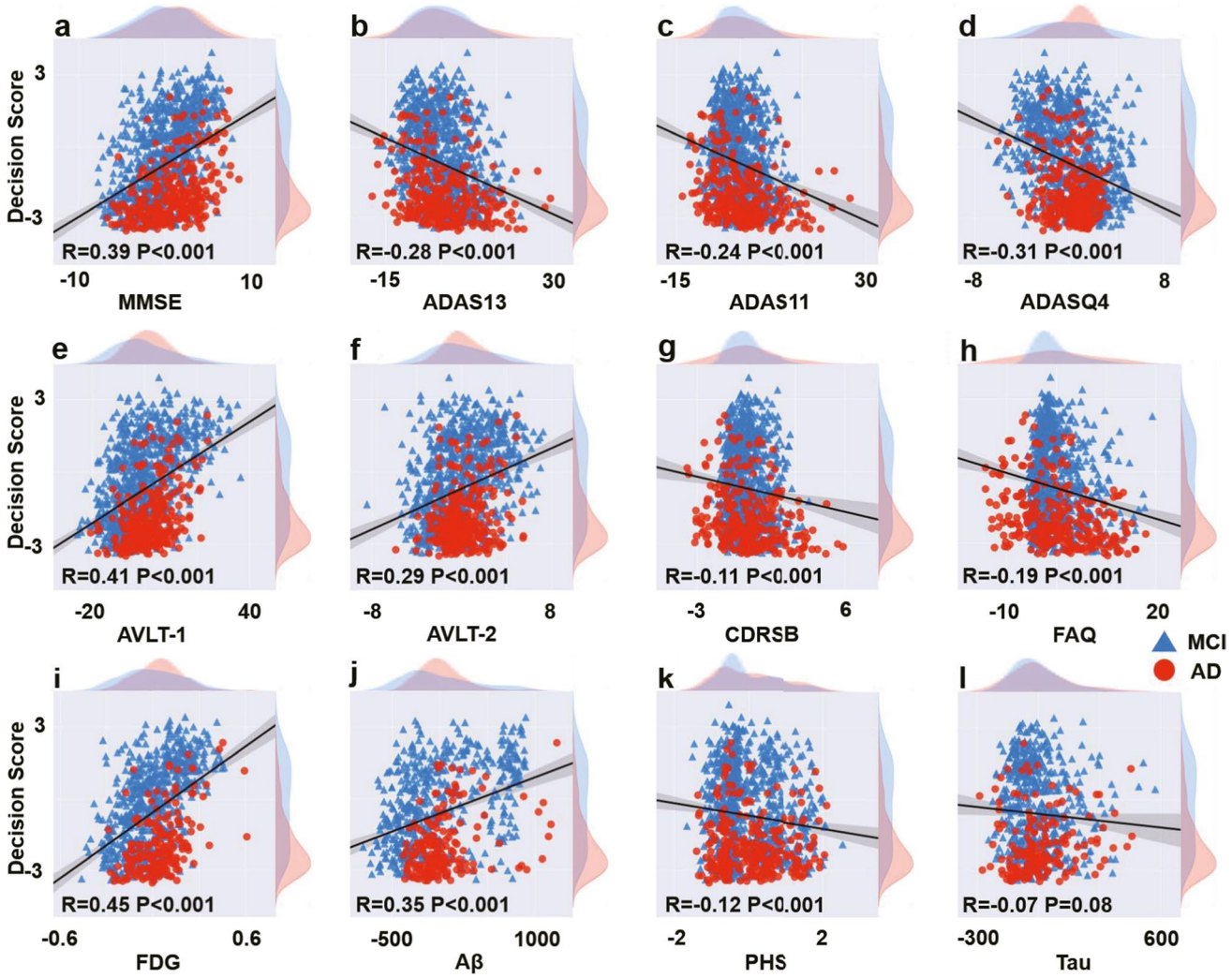

Fig. 5 Statistical analysis results of correlations between decision score and clinical profiles in the MCI and AD groups in the ADNI cohort. The clinical measures included MMSE (**a**), ADAS13 (**b**), ADAS11 (**c**), ADASQ4 (**d**), Ravlt-immediate (**e**), Ravlt-learning (**f**), CDRSB (**g**), FAQ (**h**), FDG (**i**), CSF Aβ (**j**), PHS (**k**), and CSF Tau (**l**). Note: The values of the clinical measures were plotted after regressing out the effects of age, gender, and clinical group. The decision scores of all subjects from ADNI were obtained by the classification model of AD and NC when ADNI as the testing set, and AIBL, EDSD, and OASIS as the training set

decision scores was observed among three MCI subgroups divided by CSF Aβ and CSF Tau, further revealing the solid biological substrate of the proposed biomarker. On the other hand, the decision score was significantly correlated with cognition, Aβ, FDG, and PHS, except for CSF Tau. It is well accepted that the Aβ is a specific biomarker for AD, rather than Tau. Previous studies also suggested that the isolated Tau pathway cannot trigger neurodegeneration in the brain, which might not be significantly associated with hippocampal features. Thus, this study also supports the point that the Tau deposition alone rarely leads to dementia without coexisting pathology [43].

Dynamic changes of biomarkers with the disease progressing are vitally important for monitoring the natural progression of AD. The identified alteration pattern of the decision score was highly consistent with that of neuropsychological evaluation, highlighting the sensitive tracking of disease progression, and the feasibility of being a similarity metric of abnormal hippocampal patterns concerning the AD neurodegeneration [44]. Moreover, as for the reflected dementia-like patterns, the decision score may be useful to stratify the staging along the AD spectrum to capture a better time window for severity-specific treatments [25, 44]. Thus, this work provided a solid foundation for translating neuroimaging into individual precise medicine.

Working with multiple complicated modalities indeed enhances the diagnostic performance for AD (e.g., PET scans and CSF measures) [9, 45–48], which however made

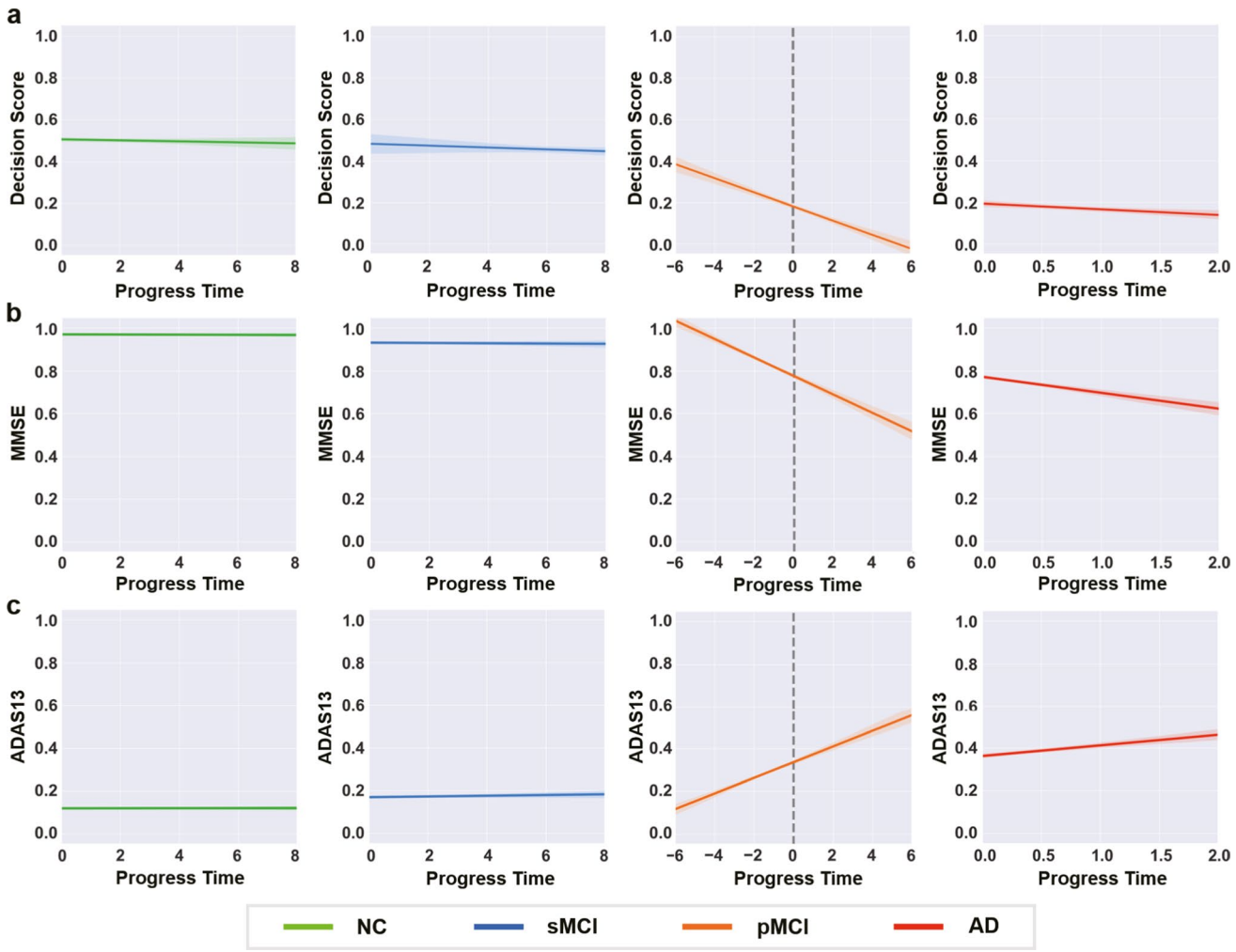

Fig. 6 Longitudinal trajectory analysis of different measures during the AD progression in the ADNI cohort. **a** Progress trajectory of the proposed decision score. **b** Progress trajectory of the MMSE score. **c** Progress trajectory of the ADAS13 score. Note: The progress time for each group was strictly aligned, and one time point was joined at an interval of 12 months if the data of this follow-up for one subject existed. The values of the decision score, MMSE, and ADAS13 scores were normalized by a max–min standardized method across all subjects. The gray dashed line in the trajectory of pMCI subjects represents the AD onset time point

Table 5 Contribution of the decision score to the overall diagnosis and prediction of AD combining with other easily-accessible clinical indicators evaluated by a linear support vector machine (SVM). Strategy (1): the grid search algorithm was capitalized to excavate the best feature combination via the accuracy of testing set (referred as best-model-fit features). Strategy (2): all involved indicators were combined to evaluate the performance. For comparison, the decision score was removed in both strategies to assess the contribution of such score to the overall diagnostic predictive potential

	Task	Feature	ACC (%)	SEN (%)	SPE (%)	AUC
Strategy (1)	pMCI vs. sMCI Internal validation	Best-model-fit features	83.0	74.5	88.1	0.90
		Best-model-fit minus decision score	80.3	65.8	88.9	0.88
	AD vs. NC External validation	Best-model-fit features	94.8	86.9	97.6	0.97
		Best-model-fit Minus Decision score	91.5	75.1	97.8	0.96
Strategy (2)	pMCI vs. sMCI Internal validation	All 11 indicators	82.3	71.8	88.5	0.90
		All 11 indicators minus decision score	81.3	67.8	89.3	0.88
	AD vs. NC External validation	All 4 indicators	94.6	86.5	97.6	0.97
		All 4 indicators minus decision score	90.6	72.2	97.9	0.96

the data acquisition more laborious and costly, and probably cause a reduction of generalization because of certain inaccessible modalities in most of hospitals. For instance, the PET scanning is a lengthy travel with invasive and expensive, which tremendously increases patient burdens and limits its universality; the extraction of CSF is very intractable due to the invasive lumbar puncture that usually causes side effects and requires hospitalization. Of note, our work built only upon structural MRI to accurately assess AD status within a non-invasive, easily accessible, and cost-effective manner [49–51]. Nevertheless, it is of interest in future studies to determine whether the currently presented well-defined patterns could be identified from the other neuroimaging scans.

There are several limitations worthy of being considered in present study. First, the involved public materials were centrally selected, yet the effectiveness and generalizability of our method should be further verified on "real-world" data for higher clinical applicability. Second, the performance was greatly influenced by the unbalanced data in each cohort where the NC is more than AD, which may bring on the poor sensitivity to an extent, although we also devised a threshold strategy to address such imbalance. Third, despite the versatility of T1 MRI and its verified performance in AD analysis, one simple MR imaging modality is still limited. Fourth, whether the comprehensive hippocampal characterization could stand the position of biomarker for AD in presence of other neurodegenerative diseases needs to be further investigated in a more diverse data.

Collectively, we developed a neuroimaging biomarker to delineate the neurodegeneration for AD with a comprehensive characterization of hippocampal feature ensemble, including gray matter volume, probability matrix, and radiomics features. Further, we validated its strong generalization, solid biological basis, and dynamic longitudinal alterations based on 3238 participants from ADNI, AIBL, EDSD, and OASIS cohorts. Study findings suggest that the comprehensive characterization of hippocampal feature ensemble is better to provide an individualized, generalizable, and biologically plausible neuroimaging biomarker for AD with promising prospects for clinical applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-023-09519-x.

Acknowledgements Data collection and sharing for this project was funded by four independent databases of Alzheimer's Disease Neuroimaging Initiative (ADNI), Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL), the European DTI Study on Dementia (EDSD), and the Open Access Series of Imaging Studies (OASIS).

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

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 of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Hongming Li, the second author, University of Pennsylvania) has significant statistical expertise and rich experience in biological research.

Informed consent Informed written consent was obtained from all participants.

Ethical 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 verify whether a comprehensive characterization of hippocampal features of gray matter volume, segmentation probability, and radiomics features could better distinguish AD from NC, and to investigate whether the classification decision score could serve as a robust and individualized brain signature.

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
- diagnostic or prognostic study
- multicenter study

References

1. Querfurth HW, Laferla FM (2010) Alzheimer's disease. N Engl J Med 362:329
2. Petersen RC (2011) Mild cognitive impairment. N Engl J Med 364:2227–2234
3. Scheltens P, De Strooper B, Kivipelto M et al (2021) Alzheimer's disease. Lancet 397:1577–1590
4. Ezzati A, Katz MJ, Zammit AR et al (2016) Differential association of left and right hippocampal volumes with verbal episodic and spatial memory in older adults. Neuropsychologia 93:380–385
5. Wen J, Thibeau-Sutre E, Diaz-Melo M et al (2020) Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 63:101694
6. Park HY, Suh CH, Heo H, Shim WH, Kim SJ (2022) Diagnostic performance of hippocampal volumetry in Alzheimer's disease or mild cognitive impairment: a meta-analysis. Eur Radiol 32:6979–6991
7. Olsen RK, Moses SN, Riggs L, Ryan JD (2012) The hippocampus supports multiple cognitive processes through relational binding and comparison. Front Hum Neurosci 6:146

8. Lin W, Tong T, Gao Q et al (2018) Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci 12:777
9. Qiu S, Joshi PS, Miller MI et al (2020) Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain 143:1920–1933
10. Jin D, Zhou B, Han Y et al (2020) Generalizable, Reproducible, and neuroscientifically interpretable imaging biomarkers for Alzheimer's disease. Adv Sci 7:2000675
11. Jack CR Jr, Barkhof F, Bernstein MA et al (2011) Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer's disease. Alzheimers Dement 7(474–485):e474
12. 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
13. Li H, Habes M, Wolk DA, Fan Y, AsDN I (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
14. 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
15. Zhang Y, Zheng Q, Zhao K et al (2021) Early diagnosis of alzheimer's disease using 3D residual attention network based on hippocampal multi-indices feature fusion. Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III 4. Springer, pp 449–457
16. Gaser C, Dahnke R, Thompson PM, Kurth F, Luders E (2022) CAT-a computational anatomy toolbox for the analysis of structural MRI data. BioRxiv:2022.2006. 2011.495736
17. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011) A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54:2033–2044
18. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, ICCV 2015, pp 1026–1034
19. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:14126980
20. Green RC, Roberts JS, Cupples LA et al (2009) Disclosure of APOE genotype for risk of Alzheimer's disease. N Engl J Med 361:245–254
21. Seshadri S, Fitzpatrick AL, Ikram MA et al (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA 303:1832–1840
22. Ashford MT, Veitch DP, Neuhaus J, Nosheny RL, Tosun D, Weiner MW (2021) The search for a convenient procedure to detect one of the earliest signs of Alzheimer's disease: a systematic review of the prediction of brain amyloid status. Alzheimers Dement 17:866–887
23. Schindler SE, Gray JD, Gordon BA et al (2018) Cerebrospinal fluid biomarkers measured by Elecsys assays compared to amyloid imaging. Alzheimers Dement 14:1460–1469
24. Shaw LM, Waligorska T, Fields L et al (2018) Derivation of cutoffs for the Elecsys amyloid β (1–42) assay in Alzheimer's disease. Alzheimer's Dement 10:698–705
25. Zhuo J, Zhang Y, Liu Y et al (2021) New trajectory of clinical and biomarker changes in sporadic Alzheimer's disease. Cereb Cortex 31:3363–3373
26. Palmqvist S, Tideman P, Cullen N et al (2021) Prediction of future Alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. Nat Med 27:1034–1042
27. Liu M, Li F, Yan H et al (2020) A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. Neuroimage 208:116459
28. Li F, Liu M, AsDN I (2019) A hybrid convolutional and recurrent neural network for hippocampus analysis in Alzheimer's disease. J Neurosci Methods 323:108–118
29. 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, CVPR 2016, pp 770–778
30. Zhao K, Ding Y, Wang P et al (2017) Early classification of Alzheimer's disease using hippocampal texture from structural MRI. In: Medical imaging 2017: biomedical applications in molecular, structural, and functional imaging, SPIE 2017, pp 625–631
31. Beheshti I, Demirel H, AsDN I (2016) Feature-ranking-based Alzheimer's disease classification from structural MRI. Magn Reson Imaging 34:252–263
32. Egan MF, Kost J, Tariot PN et al (2018) Randomized trial of verubecestat for mild-to-moderate Alzheimer's disease. N Engl J Med 378:1691–1703
33. Honig LS, Vellas B, Woodward M et al (2018) Trial of solanezumab for mild dementia due to Alzheimer's disease. N Engl J Med 378:321–330
34. Ding Y, Zhao K, Che T et al (2021) Quantitative radiomic features as new biomarkers for Alzheimer's disease: an amyloid PET study. Cereb Cortex 31:3950–3961
35. Zhao K, Zheng Q, Dyrba M et al (2022) Regional radiomics similarity networks reveal distinct subtypes and abnormality patterns in mild cognitive impairment. Advanced Science 9:2104538
36. Zhao K, Zheng Q, Che T et al (2021) Regional radiomics similarity networks (R2SNs) in the human brain: reproducibility, small-world properties and a biological basis. Netw Neurosci 5:783–797
37. Zhao K, Lin J, Dyrba M et al (2022) Coupling of the spatial distributions between sMRI and PET reveals the progression of Alzheimer's disease. Netw Neurosci 1–26
38. Woo C-W, Chang LJ, Lindquist MA, Wager TD (2017) Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 20:365–377
39. Varoquaux G (2018) Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180:68–77
40. Reiss AB, Arain HA, Stecker MM, Siegart NM, Kasselman LJ (2018) Amyloid toxicity in Alzheimer's disease. Rev Neurosci 29:613–627
41. Jack CR Jr, Knopman DS, Jagust WJ et al (2013) Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 12:207–216
42. Nussbaum JM, Schilling S, Cynis H et al (2012) Prion-like behaviour and tau-dependent cytotoxicity of pyroglutamylated amyloid-β. Nature 485:651–655
43. Suárez-Calvet M, Karikari TK, Ashton NJ et al (2020) Novel tau biomarkers phosphorylated at T181, T217 or T231 rise in the initial stages of the preclinical Alzheimer's continuum when only subtle changes in Aβ pathology are detected. EMBO Mol Med 12:e12921
44. Popuri K, Ma D, Wang L, Beg MF (2020) Using machine learning to quantify structural MRI neurodegeneration patterns of Alzheimer's disease into dementia score: Independent validation on 8,834 images from ADNI, AIBL, OASIS, and MIRIAD databases. Hum Brain Mapp 41:4127–4147
45. Spasov S, Passamonti L, Duggento A, Lio P, Toschi N, AsDN I (2019) A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. Neuroimage 189:276–287
46. El-Sappagh S, Abuhmed T, Islam SR, Kwak KS (2020) Multimodal multitask deep learning model for Alzheimer's disease

progression detection based on time series data. Neurocomputing 412:197–215
47. Liu M, Cheng D, Wang K, Wang Y (2018) Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. Neuroinformatics 16:295–308
48. Gupta Y, Lama RK, Kwon G-R et al (2019) Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. Front Comput Neurosci 13:72
49. Hansson O (2021) Biomarkers for neurodegenerative diseases. Nat Med 27:954–963
50. Chandra A, Dervenoulas G, Politis M (2019) Magnetic resonance imaging in Alzheimer's disease and mild cognitive impairment. J Neurol 266:1293–1302
51. Lehéricy S, Marjanska M, Mesrob L, Sarazin M, Kinkingnehun S (2007) Magnetic resonance imaging of Alzheimer's disease. Eur Radiol 17:347–362