Ensemble Machine Learning and SHAP-Based Analysis of Hippocampal MRI Radiomics for Classification of Cognitive Impairment

Jayden Bang[1]*

Received May 30, 2026
Accepted October 30, 2026
Electronic access December 15, 2026

Alzheimer's disease (AD) is an irreversible cause of cognitive decline, so noninvasive quantification of a patient's status is important. Especially, quantification of diagnosing AD in a non-invasive way is crucial. However, the current guideline to diagnose AD primarily relies on the fixed and written examination (i.e., MMSE) which lacks in quantifying the characteristics of human brain anatomy. In this study, we hypothesized that radiomics features extracted from the hippocampal regions on T1-weighted MRI can be used to classify cognitive status using machine learning. We showed that the ensemble of three machine learning models using hippocampal radiomics features enables diagnostic-group classification of cognitive impairment (CI) vs. cognitively normal (CN) with an AUROC of 0.787 (95% CI: 0.616–0.923), and AD vs. MCI vs. CN with a macro AUROC of 0.709 (95% CI: 0.572–0.810). We also computed SHAP values for the selected radiomics features to attribute predictions to individual features. Results imply that hippocampal radiomics may help quantify patient status to support diagnostic-group classification. All performance estimates were obtained under nested cross-validation with subject-level bootstrap confidence intervals. These findings should be interpreted with caution given the small sample and the absence of an external cohort, which limit the stability and generalizability of the reported estimates.

Introduction

Alzheimer's disease (AD) is the most common form of dementia—a group of disorders characterized by progressive decline in memory, thinking, and behavior[1]. It affects millions worldwide, particularly among the elderly, and currently, there is no known cure[1]. Thus, early and accurate detection of Alzheimer's disease is critical because the known treatments and therapies are most effective when started at an early stage, before significant brain damage occurs[1].

One of the earliest regions affected by AD is the hippocampus, a small, curved structure within the temporal lobe within the brain that plays a crucial role in memory formation and learning[2]. Studies have shown that hippocampal shrinkage, damage or other structural changes can occur several years before clinical symptoms of Alzheimer's become noticeable[3–5], making it an ideal target for diagnostic-group classification[6]. Volumetric changes in hippocampal subfields are detectable in preclinical stages, prior to subjective cognitive complaints[4] and measures of neurodegeneration in cognitively unimpaired older adults used to track subsequent decline in verbal learning and memory[5].

In recent years, a new approach called radiomics has become increasingly valuable in medical imaging research[7–9]. Radiomics analysis involves extracting large numbers of quantitative features such as shape, intensity, and texture from medical images (e.g., magnetic resonance imaging (MRI) scans). These radiomics features are extracted based on the rule-based algorithms from the target images, thus often invisible to human eyes. These features can reveal subtle differences between healthy and diseased tissues.

To interpret these complex radiomics data, researchers frequently use machine learning, a branch of artificial intelligence that teaches machines to recognize patterns from data. Machine learning algorithms, such as random forests or support vector machines (SVM), can be trained to analyze radiomics features to predict the presence or severity of diseases.

Despite this potential, relatively few studies have specifically applied hippocampal radiomics combined with machine learning methods for diagnostic-group classification and differentiation of AD from other cognitive conditions, such as mild cognitive impairment (MCI), which often precedes full AD[10–13]. While existing models demonstrate the possibility of machine learning-based radiomics for AD diagnosis[14,15], their transportability is constrained in two ways: radiomic feature values are highly sensitive to differing hippocampal segmentation method used[16] and to differing scanner and acquisition settings[17,18]. Further, model outputs are rarely bolstered by feature-level explanation of which radiomic properties drove predictions[19].

In this study, we aim to investigate whether radiomics features extracted from hippocampal T1-weighted MRI scans,

[1] *Korea International School Pangyo; Seongnam, South Korea.*
* *Corresponding author: jayden3bang@gmail.com*

© The National High School Journal of Science 2026

combined with ensemble of machine learning methods, could accurately differentiate among AD, MCI, and cognitively normal controls (CN). Our goal is to develop a reproducible, explainable, and non-invasive diagnostic tool that could support early clinical decision-making and ultimately improve patient care outcomes.

Contributions

- Developed an ensemble machine learning framework (Random Forest, XGBoost, and AdaBoost) combined with hippocampal radiomics features extracted from T1-weighted MRI to perform diagnostic-group classification of CN, MCI, and AD.
- Reported nested cross-validated performance with bootstrap confidence intervals: a hippocampal radiomics-only AUROC of 0.787 (95% CI: 0.616–0.923) for the CI vs. CN task and a macro AUROC of 0.709 (95% CI: 0.572–0.810) for the CN/MCI/AD multiclass task.
- Applied SHAP analysis to quantify how individual radiomic features contribute to model predictions, supporting interpretation of the model's outputs.
- Released the code and extracted radiomic features publicly to support reproducibility of hippocampal radiomics research.

Methods and Materials

Dataset

This study utilized a dataset containing T1-weighted magnetic resonance imaging (MRI) scans from a total of 135 participants from the ADNI cohort. Each MRI scan included segmented masks clearly identifying the left and right hippocampal regions. The participants were categorized into three diagnostic groups: [1] Cognitively Normal (CN): 44 patients, [2] Mild Cognitive Impairment (MCI): 46 patients [3] Alzheimer's Disease (AD): 45 patients.

Instead of being separated into fixed training and test sets, all 135 subjects were analyzed within a single nested cross-validation design. In each repetition, the cohort was partitioned into stratified subject-level folds. The model was then built on the training folds of a split and evaluated on the fold held out from that same split, so that every subject served as evaluation data exactly once per repetition. Because all folds are drawn from the same ADNI-derived sample rather than from a separate cohort, scanner set, institution, or ADNI phase, no partition used here constitutes an independent validation cohort.

To check for demographic imbalance between the groups, we compared group demographics prior to modeling. The mean age was 76.2 $\pm$ 7.5 years in the CN group ($n = 44$; range 61–90), 74.7 $\pm$ 8.1 years in the MCI group ($n = 46$; range 60–87), and 74.4 $\pm$ 8.1 years in the AD group ($n = 45$; range 63–90). The sex distribution was 22 male / 22 female (CN), 27 male / 19 female (MCI), and 21 male / 24 female (AD). Within the three diagnostic groups, there was no statistically significant difference in age (Kruskal-Wallis $H = 1.14$, $p = 0.565$) and no statistically significant difference in sex distribution (chi-square $= 1.41$, df $= 2$, $p = 0.495$), indicating the groups were balanced with respect to these variables. Since the groups were also nearly equal in size (44 / 46 / 45), no resampling was applied; the Random Forest used balanced class weights.

All reported performance was obtained under nested cross-validation. Normalization, feature selection, hyperparameter search, and model selection were performed strictly within training folds, with no information from evaluation folds entering any of those steps. Evaluation used stratified subject-level folds, and the Random Forest used balanced class weights; no synthetic oversampling was applied. Confidence intervals were estimated by subject-level bootstrap resampling of the saved outer-fold predictions. With a fixed seed, the outer loop used 5-fold stratified subject-level cross-validation repeated 10 times (50 outer folds); the inner loop used 5-fold cross-validation for hyperparameter search, and feature selection was capped at 64 features per fold. Confidence intervals for the headline metrics come from 1,000 subject-level bootstrap resamples of the saved outer-fold predictions; the class-wise and paired classifier comparisons reported below use 100 resamples. Analyses were run under Python 3.8.10 on macOS.

Feature dimensionality was also quantified. From 1,702 raw radiomic features extracted across 135 subjects, feature selection retained 64 features per fold, corresponding to a selected-feature-to-sample ratio of 0.47 rather than the raw ratio of 12.6, which reduces the risk of overfitting associated with high-dimensional feature spaces.

SHAP values were computed per outer fold using TreeExplainer with interventional perturbation on raw model output where supported, falling back to a probability-based explainer otherwise, with a background set of 50 training subjects and 27 evaluation subjects per fold.

Radiomic feature extraction

For the first step for the entire pipeline for our method (Fig 1.), we first extracted radiomics features—quantitative measures describing the shape, intensity, and texture—from the segmented hippocampal regions of each MRI scan. These features enabled the identification of subtle structural differences between patients with Alzheimer's disease (AD), mild cognitive impairment (MCI), and cognitively normal (CN) controls.

We used PyRadiomics package (version 2.2) to extract the

© The National High School Journal of Science 2026

radiomics features from the left and right hippocampal regions from the T1-weighted MRI. Before feature extraction, we preprocessed the raw MRI imaging using z-score normalization using Numpy (version 1.24.4). We then applied wavelet transformation to extract wavelet radiomics features which may provide rich visionary information from the raw T1-weighted MRI images.

The categories of radiomic features extracted, and how often features from each were retained, are summarized in Table 5.

Machine learning model development

For the machine learning model, we used an ensemble voting classifier, which combines multiple machine learning models to enhance predictive accuracy[20]. Our ensemble consisted of three distinct classifiers: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and AdaBoost (Adaptive Boosting). We employed a systematic grid-search method with 5-fold cross-validation within the training folds of each outer split to choose our best hyperparameters and optimize each individual model within the ensemble. After individual optimization, the three models were combined back into a single voting classifier, which then made predictions based on the weighted average of the outputs of each model (i.e., soft voting approach). The resulting voting classifier was then evaluated on the fold held out from that split, which by design contributed no information to normalization, feature selection, or hyperparameter search.

Evaluation

We evaluated the performance of our selected machine learning model on the outer folds of the nested cross-validation; reported values are means over the 50 outer folds, with confidence intervals from subject-level bootstrap resampling of the saved outer-fold predictions. The primary metric used to measure performance was the area under the receiver operating characteristic curve (AUROC). The AUROC provides an assessment of the ability of the model to correctly classify patients, with a higher AUROC indicating better predictive accuracy.

We experimented two specific tasks: [1] Binary classification task: Differentiating between cognitive impairment (CI; MCI or AD) and cognitively normal (CN) participants, [2] Multiclass classification: Differentiating among AD, MCI, and CN patients simultaneously.

It is important to interpret the trained models to show the explainability of the model[21]. We calculated SHAP values from the selected features to show how individual features contribute to the model outputs (Figure 4)[22]. For the reproducibility of research, we release our code and extracted features at `https://github.com/JaydenBang/MRIRadiomics`.

Results and discussion

Utilizing ensemble of machine learning models (i.e., XGBoost, AdaBoost, and Random Forest), we analyzed radiomics features, specifically those of the hippocampus. Using a dataset consisting of CN/MCI/AD labeled MRI scans. In Table 1, we show that our model achieved a hippocampal radiomics-only AUROC of 0.787 (95% CI: 0.616–0.923) when differentiating between CI and CN as a binary classification task (Figure 2), and a macro AUROC of 0.709 (95% CI: 0.572–0.810) when classifying between AD, MCI, and CN groups under nested cross-validation (Figure 3). Performance varied markedly across the individual pairwise contrasts: CN vs. AD reached an AUROC of 0.879 (95% CI: 0.700–1.000), whereas CN vs. MCI reached 0.678 (95% CI: 0.481–0.910) and AD vs. MCI reached 0.632 (95% CI: 0.372–0.807). A demographics-only model using age and sex performed near chance on the same splits (CI vs. CN AUROC 0.584; three-class macro AUROC 0.493), and adding demographics to the radiomics features changed performance only marginally (CI vs. CN AUROC 0.796; three-class macro AUROC 0.712), indicating that the discriminative signal is carried by the hippocampal radiomic features rather than by demographic separation. Across classifiers, Random Forest performed best, and differences between classifiers were not significant after correction for multiple comparisons (all FDR-adjusted $p > 0.16$); soft voting did not outperform the best individual model. Per-classifier results for every task are given in Table 1, class-wise performance for the three-class task in Table 4, and a comparison with previously published hippocampal radiomics studies in Table 2.

Our small dataset may bring the risk of irregular performance which means the lack of robustness. Thus, further data collection in the form of labeled MRI data is required. For multiclass classification, our model achieved a macro AUROC of 0.709 (95% CI: 0.572–0.810), with a macro balanced accuracy of 0.533 and a macro F1 of 0.523. The width of this interval indicates considerable uncertainty at this sample size. However, the confidence and definitive diagnosis are necessary in the clinical practice for patients, meaning a higher AUROC must be acquired. The relatively low performance of multiclass classification may reflect the continuum of hippocampal atrophy across MCI and AD, where MCI occupies an intermediate stage and the structural boundary between adjacent groups is less distinct than between CI and CN[24]. This can be mitigated by integrating a multimodal data model, such as using pTAU or amyloid-beta plaque information[25].

Across the three-class model, the features most attributed by SHAP differ from class to class. While surface-to-volume

© The National High School Journal of Science 2026

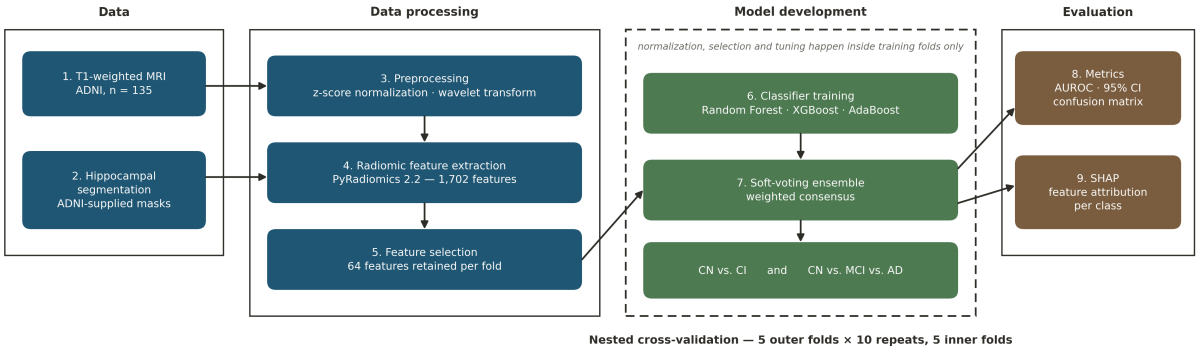

Figure. 1 Overall framework. The pipeline runs from left to right through nine steps. T1-weighted MRI and hippocampal segmentation masks are inputted then sent for preprocessing through z-score normalization and wavelet transform; radiomic feature extraction with PyRadiomics 2.2 (1,702 features); feature selection to 64 features per fold; training of Random Forest, XGBoost and AdaBoost; soft-voting consensus; and evaluation by AUROC with 95% confidence intervals, confusion matrices and SHAP attribution. The dashed border marked around model development identifies what sits inside the nested cross-validation loop (5 outer folds × 10 repeats, 5 inner folds): normalization, feature selection and hyperparameter tuning all occur within training folds, therefore no information from an evaluation fold reaches any modeling decision.

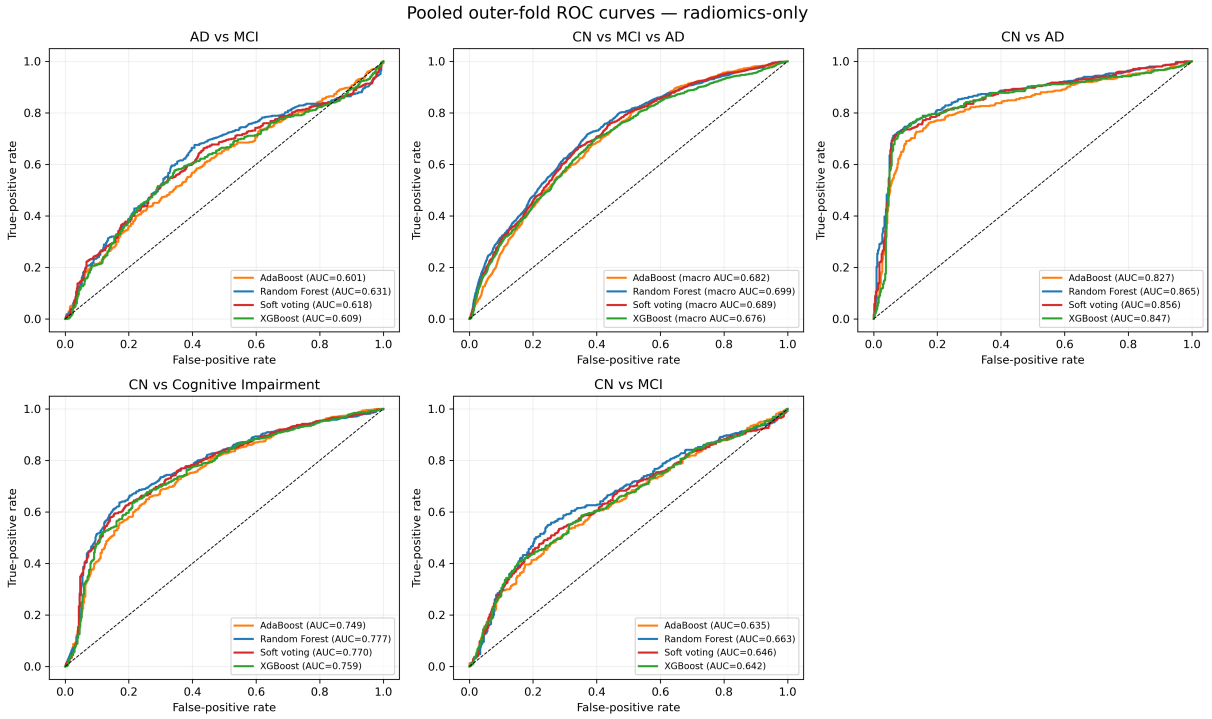

Figure. 2 Receiver operating characteristic (ROC) curves pooled across the 50 outer folds for all five diagnostic tasks and all four classifiers, hippocampal radiomics features only. Discrimination is ordered consistently across classifiers: CN vs. AD is strongest, CN vs. cognitive impairment intermediate, and the contrasts involving MCI approach the diagonal. The AUROC values printed in each panel are computed on the pooled predictions and therefore differ slightly from the fold-averaged values reported in Table 1 (for example 0.777 pooled against 0.787 fold-averaged for CN vs. CI with Random Forest), the fold-averaged values are the ones the confidence intervals belong to.

© The National High School Journal of Science 2026

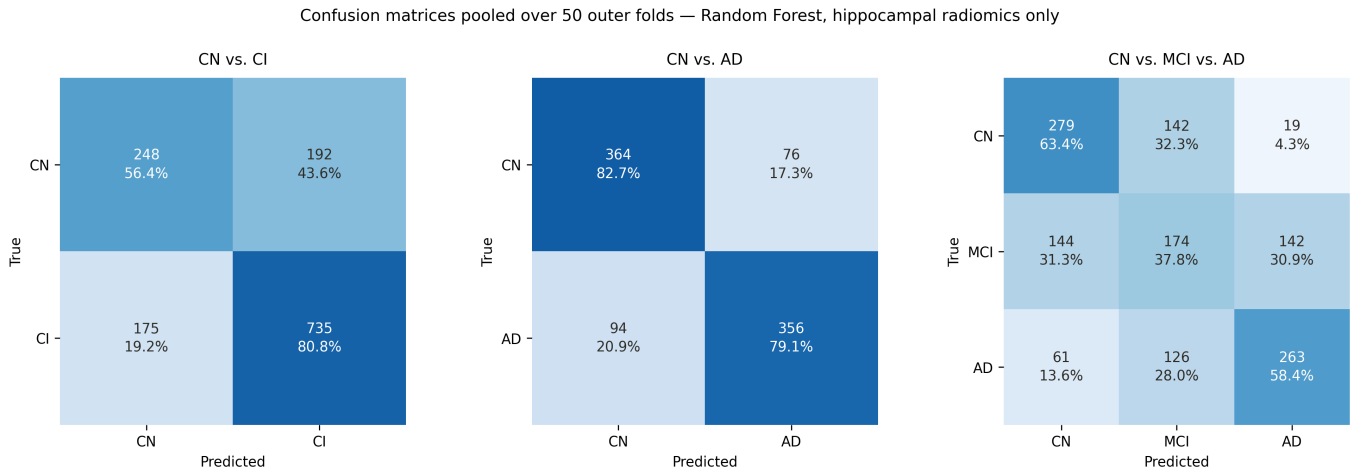

Figure. 3 Confusion matrices pooled over the 50 outer folds for the Random Forest model using hippocampal radiomics features, with counts and row-normalized percentages. The residual error is asymmetric in the binary task: 735 of 910 impaired cases are recovered (80.8%) but only 248 of 440 CN cases (56.4%), so the model errors toward calling a cognitively normal subject impaired. In the three-class task MCI is the limiting class, correctly assigned in 174 of 460 cases (37.8%) and split almost evenly between the CN and AD columns, which is the structural overlap the continuum interpretation predicts.

Table 1 Consolidated model performance, hippocampal radiomics features only. All values are means over 50 outer folds (5-fold cross-validation repeated 10 times) under nested cross-validation, with 95% confidence intervals from 1,000 subject-level bootstrap resamples of the saved outer-fold predictions. Precision, recall, specificity and F1 are macro-averaged. For the three-class task the AUROC is the macro one-vs-rest value. Random Forest gives the highest point estimate in every task, but no pairwise difference between classifiers survives correction for multiple comparisons (minimum FDR-adjusted $p = 0.16$), and soft voting does not outperform the best individual classifier.

Task	Classifier	Acc.	Bal. acc.	Prec.	Rec.	Spec.	F1	AUROC (95% CI)
CI vs. CN	Random Forest	0.728	0.686	0.701	0.686	0.564	0.681	0.787 (0.616–0.923)
	XGBoost	0.711	0.667	0.676	0.667	0.541	0.663	0.764 (0.582–0.887)
	AdaBoost	0.721	0.670	0.689	0.670	0.523	0.669	0.758 (0.553–0.874)
	Soft voting	0.726	0.681	0.695	0.681	0.550	0.678	0.780 (0.600–0.900)
CN vs. AD	Random Forest	0.809	0.809	0.822	0.809	0.827	0.806	0.879 (0.700–1.000)
	XGBoost	0.810	0.810	0.824	0.810	0.831	0.807	0.865 (0.706–0.988)
	AdaBoost	0.789	0.789	0.801	0.789	0.811	0.784	0.861 (0.685–1.000)
	Soft voting	0.802	0.801	0.816	0.801	0.821	0.798	0.874 (0.690–1.000)
CN vs. MCI	Random Forest	0.628	0.629	0.639	0.629	0.632	0.620	0.678 (0.481–0.910)
	XGBoost	0.606	0.606	0.615	0.606	0.614	0.598	0.654 (0.443–0.852)
	AdaBoost	0.603	0.604	0.610	0.604	0.607	0.595	0.651 (0.457–0.823)
	Soft voting	0.601	0.602	0.610	0.602	0.612	0.594	0.660 (0.472–0.861)
AD vs. MCI	Random Forest	0.627	0.627	0.639	0.627	0.654	0.619	0.632 (0.372–0.807)
	XGBoost	0.606	0.606	0.619	0.606	0.613	0.597	0.608 (0.336–0.787)
	AdaBoost	0.583	0.582	0.589	0.582	0.594	0.576	0.604 (0.379–0.766)
	Soft voting	0.607	0.607	0.614	0.607	0.618	0.600	0.624 (0.373–0.753)
CN vs. MCI vs. AD	Random Forest	0.530	0.533	0.537	0.533	0.533	0.523	0.709 (0.572–0.810)
	XGBoost	0.500	0.502	0.506	0.502	0.502	0.492	0.680 (0.541–0.815)
	AdaBoost	0.504	0.505	0.531	0.505	0.505	0.503	0.692 (0.543–0.832)
	Soft voting	0.516	0.518	0.529	0.518	0.518	0.513	0.699 (0.560–0.838)

© The National High School Journal of Science 2026

Table 2 Comparison with previously published hippocampal radiomics studies. Reported AUROC values are not directly comparable across rows, because cohort composition, task definition, and validation strategy differ; entries marked "not reported" were not specified in the source publication. Two features of this comparison are relevant to the present work. First, the studies reporting the highest AUROC values generally evaluate on a single train/test partition, whereas the present estimates are obtained under nested cross-validation with confidence intervals, which is the more conservative procedure. Second, the larger studies achieve their performance with external validation cohorts that were not available here, which is the principal limitation of the present work rather than the modeling approach.

Study	Cohort (n)	Segmentation	Features	Classifier	Validation	AUROC
Feng et al., 2019 [15]	Single centre, 86 (42 aMCI / 44 NC)	Learning-based ROI with deformable shape prior	385 (histogram, GLCM, Haralick, RLM)	Logistic regression (per side)	70/30 split, 10 repetitions	0.79 left, 0.76 right (aMCI vs. NC)
Du et al., 2022 [23]	ADNI, 144 (36 EOAD / 36 LOAD / 36 young / 36 old controls) + 60 external	Atlas-based (AAL) with manual inspection, 3D-Slicer	214, PyRadiomics	Support vector machine (RBF)	70/30 split + external cohort	0.90 EOAD vs. young; 0.94 LOAD vs. old; 0.87 EOAD vs. LOAD
Wang et al., 2022 [14]	Single centre, 178 (84 AD / 50 aMCI / 44 NC) + 99 ADNI external	Hippodeep (deep learning)	101 per modality (13 intensity, 88 texture), IBSI-conforming	LASSO radiomics score	70/30 split x 10 + ADNI external	0.818 AD vs. NC; 0.713 aMCI vs. NC (structural MRI, left)
Xia et al., 2024 [10]	729 across four cohorts (ADNI, OASIS, clinical centre)	Hippodeep (deep learning)	214, PyRadiomics 3.0.1	Logistic regression (12 features)	10-fold CV, 7:3 split, three external cohorts	0.91 training; 0.91 internal; 0.93 OASIS; 0.87 clinical
Yin et al., 2024 [11]	Single centre, 234 (126 AD / 108 HC)	Not reported	1,037 per side (11 retained)	Random Forest, logistic regression, SVM	Not reported	0.961 AD vs. HC (RF); 0.875 AD-MCI vs. AD-D
This study	ADNI, 135 (44 CN / 46 MCI / 45 AD)	ADNI-supplied hippocampal masks	1,702 (PyRadiomics 2.2, including wavelet) ; 64 retained per fold	Random Forest, XGBoost, AdaBoost, soft voting	Nested cross-validation with subject-level bootstrap 95% CIs	0.787 (0.616–0.923) CI vs. CN; 0.879 (0.700–1.000) CN vs. AD; 0.709 (0.572–0.810) three-class

Table 3 Class-wise diagnostic performance for the three-class task (CN vs. MCI vs. AD), hippocampal radiomics features with the Random Forest classifier. Values are computed from the pooled outer-fold predictions; intervals are 95% percentile intervals from 100 subject-level bootstrap resamples. MCI is the limiting class on every metric, with a sensitivity of 0.378 against 0.634 for CN and 0.584 for AD, which is the main source of the modest three-class performance.

Class	Sens. (95% CI)	Spec. (95% CI)	Prec. (95% CI)	F1 (95% CI)
CN	0.634 (0.501–0.752)	0.775 (0.659–0.826)	0.576 (0.412–0.673)	0.604 (0.472–0.689)
MCI	0.378 (0.282–0.508)	0.699 (0.634–0.768)	0.394 (0.263–0.538)	0.386 (0.280–0.490)
AD	0.584 (0.458–0.686)	0.821 (0.752–0.882)	0.620 (0.472–0.731)	0.602 (0.483–0.688)

ratio of the left hippocampus ranks fourth for CN and ninth for MCI, for AD it remains below the top twenty-five. Gray-level run-length and dependence features account for 23 of the 25 highest-attribution features for AD, against 19 for CN (Figure 5).

While we do not test the association directly, texture-based hippocampal features have previously been reported as discriminative for AD and amnestic MCI relative to controls using comparable feature families [11,15,23], and our attribution from SHAP is consistent with those reports. It is important to note that feature selection favored the left hippocampus in all classes (24 of the 25 highest-attribution features, 52.3 of 64 retained features per fold overall), which parallels earlier work reporting stronger discrimination from the left hemisphere [14,15]. These attribution values are drawn from a single outer fold and should be read as indicative rather than stable.

To improve the clinical integration of our proposed method, the classification of CI can be achieved prior to the segmentation of hippocampus regions from T1-weighted MRI scans. This is not trivial due to the complex anatomy of the hippocampal area and high-variability of the segmentation performance in medical imaging [26]. Moreover, we used the public datasets of ADNI with the contoured hippocampus regions, which have limited number of datasets without the scalability. At last, the radiomics analysis greatly relies on the quality of the target segmentation masks, it is required to make the contours clear and robust before extracting and analyzing the extracted features to make the entire process reliable and reproducible. Radiomic feature values are also sensitive to acquisition: field strength, vendor, voxel size and reconstruction all shift texture and first-order values, and post-extraction harmonization such as ComBat has been shown to recover feature reproducibility and classification performance across scanners and extraction settings [18,27].

In this study, hippocampal radiomic features were evaluated for their ability to distinguish cognitive diagnostic groups

© The National High School Journal of Science 2026

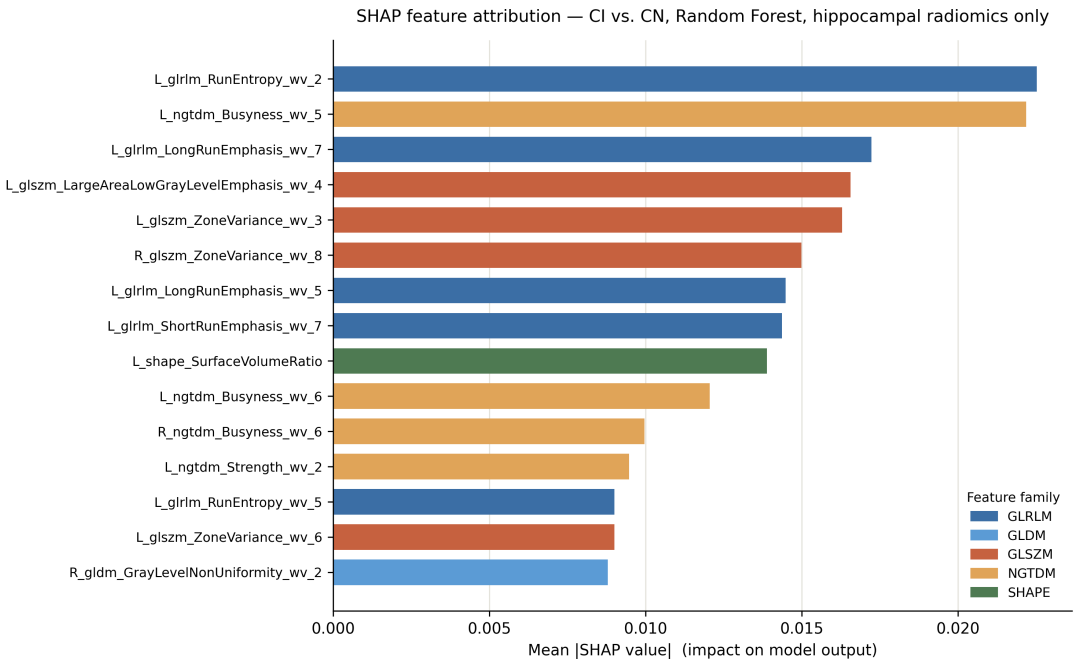

Figure. 4 SHAP feature attribution for the CI vs. CN Random Forest model, hippocampal radiomics features only, showing the fifteen features with the largest mean absolute SHAP value and colored by feature family. Because this is the radiomics-only model, age and sex are not inputs and cannot appear. The ranking is dominated by wavelet-derived texture rather than by shape: the three highest-attribution features are a GLRLM run-entropy, an NGTDM busyness and a GLRLM long-run-emphasis feature, all from the left hippocampus and all wavelet-derived, and only one shape feature (left surface-to-volume ratio) enters the top ten. SHAP values were computed per outer fold with TreeExplainer using interventional perturbation on raw model output, with a background set of 50 training subjects and 27 evaluation subjects per fold; the values plotted are from the first outer fold.

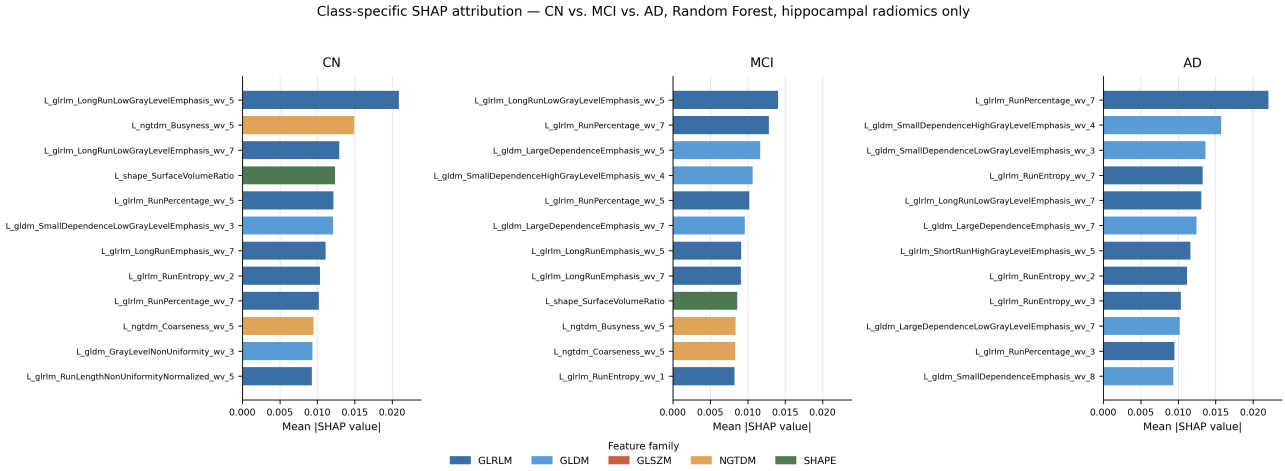

Figure. 5 Class-specific SHAP attribution for the three-class task (CN vs. MCI vs. AD), Random Forest, hippocampal radiomics features only, showing the twelve features with the largest mean absolute SHAP value for each class. Attribution differs by class: surface-to-volume ratio of the left hippocampus is the fourth-ranked feature for CN and ninth for MCI but does not enter the top twenty-five for AD, where gray-level run-length and dependence features dominate. The left hippocampus supplies 24 of the 25 highest-attribution features in every class. Values are computed on raw model output with TreeExplainer and are drawn from the first outer fold; the ranking should be read as indicative rather than stable.

© The National High School Journal of Science 2026

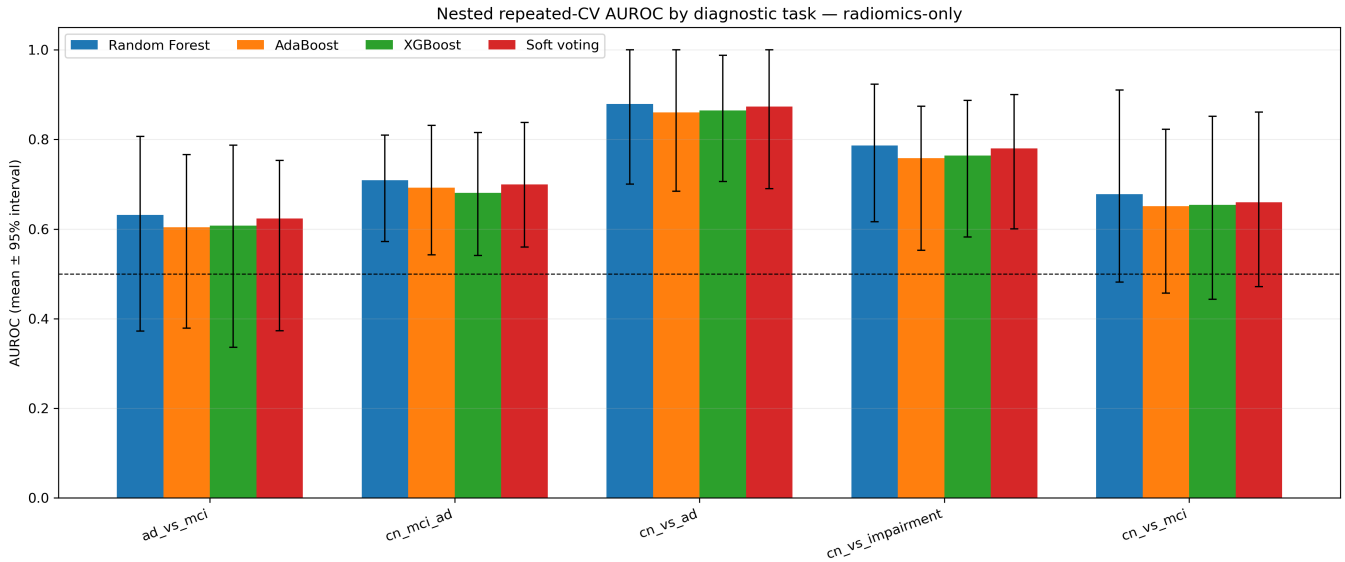

Figure. 6 Nested repeated cross-validation AUROC by diagnostic task and classifier, hippocampal radiomics features only, with 95% confidence intervals. Random Forest holds the highest point estimate in every task, but the intervals overlap heavily across classifiers and all cross the interval of the soft-voting ensemble; no pairwise difference survives correction for multiple comparisons (minimum FDR-adjusted $p = 0.16$). The intervals for the contrasts involving MCI cross 0.5, so those tasks are not distinguishable from chance at this sample size.

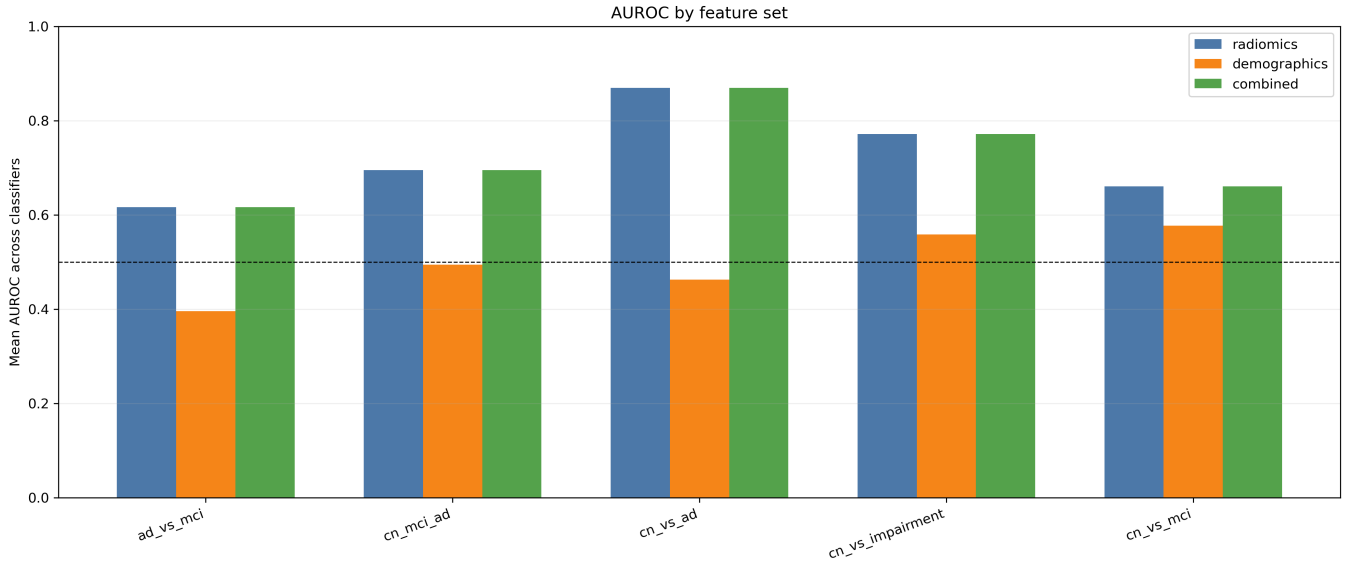

Figure. 7 Mean AUROC by feature set across the four classifiers, for each diagnostic task. A demographics-only model built from age and sex performs at or below chance on every task (0.396 for AD vs. MCI, 0.463 for CN vs. AD, 0.493 for the three-class task), while radiomics-only and combined models are close to identical everywhere. Adding age and sex to the radiomic features changes performance only marginally, which is the direct evidence that the reported discrimination is carried by hippocampal texture and shape rather than by demographic separation between the diagnostic groups.

in a 135-subject ADNI dataset. Using repeated nested cross-validation with subject-level bootstrap confidence intervals, the random forest model achieved an AUROC of 0.879 (95% CI: 0.700–1.000) for CN versus AD, 0.678 (95% CI: 0.481–

© The National High School Journal of Science 2026

Table 4 Categories of radiomic features extracted from the left and right hippocampus and how often each was retained. A total of 1,702 features were extracted per subject (851 left, 851 right) and exactly 64 were retained per fold; the family counts are means over the 50 outer folds of the CI vs. CN Random Forest model and sum to 64. Two observations are worth noting: no first-order feature was retained in any fold, so the signal is carried entirely by shape and texture families; and 91.5% of all retained features are wavelet-derived, a property that cuts across the families rather than forming one of them. Selection also favored the left hippocampus (52.3 of 64 features per fold on average, against 11.7 from the right), and 195 distinct features were selected at least once.

Feature family	What it quantifies	Mean retained per fold (of 64)
First-order / histogram	Distribution of voxel intensities inside the hippocampal mask (mean, variance, skewness, kurtosis, energy)	0 (none retained)
Shape	Geometry of the segmented hippocampus, independent of intensity (volume, surface area, sphericity, elongation)	5.2
GLRLM (gray level run length matrix)	Length of consecutive voxels sharing an intensity value	24.5
GLSZM (gray level size zone matrix)	Size of connected zones of equal intensity	5.6
GLDM (gray level dependence matrix)	Dependence of a voxel's intensity on its immediate neighbours	23.2
NGTDM (neighbouring gray tone difference matrix)	Contrast and busyness between a voxel and its surrounding neighbourhood	5.5
Wavelet-derived	Any of the above computed after wavelet decomposition of the image	91.5% of retained features (cross-cutting)

0.910) for CN versus MCI, 0.632 (95% CI: 0.372–0.807) for AD versus MCI, and 0.709 (95% CI: 0.572–0.810) for the three-class classification task. For the binary cognitive impairment versus cognitively normal task, the AUROC was 0.787 (95% CI: 0.616–0.923). To check for demographic bias, a demographics-only model was tested, which performed near chance on the same splits (AUROC 0.584), suggesting that the discriminative signal is carried not by age or sex, but by the imaging features alone. Our findings suggest that hippocampal radiomics may have the ability to capture differences between established diagnostic groups, particularly CN and AD. It is important to note, however, that this study does not suggest early Alzheimer's disease diagnosis, progression prediction, or any clinical readiness. The analyzed labels were assigned at the time of imaging, so the task our model performed is diagnostic-group classification, not early detection. The available data also did not include external cohorts, scanner information, longitudinal outcomes, biomarkers, or detailed clinical variables.

In the future, larger multi-site and longitudinal studies should evaluate whether hippocampal radiomics provides additional value beyond routine clinical information. External validation on cohorts acquired outside ADNI, testing of feature stability across scanners and segmentation operators, separation of stable from progressive MCI with longitudinal follow-up[28], biomarker comparisons, and clinical workflow evaluation will be necessary before this approach can be considered for clinical decision support. Within those limits, these results support hippocampal radiomics as a quantitative adjunct worth further investigation rather than as a diagnostic tool.

Table 5 Abbreviations and nomenclature. Note that CI denotes cognitive impairment when used as a class label and confidence interval when used with an interval estimate; the two are distinguished by context and by the "95%" prefix.

Abbreviation	Definition
AD	Alzheimer's disease
MCI	Mild cognitive impairment
aMCI	Amnestic mild cognitive impairment
CN	Cognitively normal
CI	Cognitive impairment (MCI or AD), when used as a class label
95% CI	95% confidence interval, when used with an interval estimate
ADNI	Alzheimer's Disease Neuroimaging Initiative
MRI	Magnetic resonance imaging
T1WI	T1-weighted imaging
ROC	Receiver operating characteristic
AUROC	Area under the receiver operating characteristic curve
RF	Random Forest
XGBoost	Extreme Gradient Boosting
AdaBoost	Adaptive Boosting
SVM	Support vector machine
CV	Cross-validation
FDR	False discovery rate
SHAP	Shapley additive explanations
GLCM	Gray level co-occurrence matrix
GLRLM	Gray level run length matrix
GLSZM	Gray level size zone matrix
GLDM	Gray level dependence matrix
NGTDM	Neighbouring gray tone difference matrix
IBSI	Image Biomarker Standardization Initiative
MMSE	Mini-Mental State Examination
pTAU	Phosphorylated tau

References

1 A. Association, 2024 alzheimer's disease facts and figures. Alzheimer's & Dementia. Vol. 20, pg. 3708–3821, 2024., https://doi.org/10.1002/alz.13809.

2 F. D. Raslau, I. T. Mark, A. P. Klein, J. L. Ulmer, V. Mathews, L. P. Mark, Memory part 2: the role of the medial temporal lobe. American Journal of Neuroradiology. Vol. 36, pg. 846–849, 2015., https://doi.org/10.3174/ajnr.A4169.

3 N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, M. N. Rossor, Presymptomatic hippocampal atrophy in alzheimer's diseasea longitudinal mri study. Brain. Vol. 119, pg. 2001–2007, 1996., https://doi.org/10.1093/brain/119.6.2001.

4 Z. Yildirim, F. Delen, D. Berron, H. Baumeister, G. Ziegler, H. Schütze, W. Glanz, L. Dobisch, O. Peters, S. D. Freiesleben, et al., Brain reserve contributes to distinguishing preclinical alzheimer's stages 1 and 2. Alzheimer's Research & Therapy. Vol. 15, pg. 43, 2023., https://doi.org/10.1186/s13195-023-01187-9.

© The National High School Journal of Science 2026

5 S. L. Allison, E. M. Jonaitis, R. L. Koscik, B. P. Hermann, K. D. Mueller, R. P. Cary, Y. Ma, H. A. Rowley, C. M. Carlsson, S. Asthana, H. Zetterberg, K. Blennow, B. B. Bendlin, S. C. Johnson, Neurodegeneration, alzheimer's disease biomarkers, and longitudinal verbal learning and memory performance in late middle age. Neurobiology of Aging. Vol. 102, pg. 151–160, 2021., https://doi.org/10.1016/j.neurobiolaging.2021.01.030.

6 Z. Chen, S. Bi, Y. Shan, B. Cui, H. Yang, Z. Qi, Z. Zhao, Y. Han, S. Yan, J. Lu, Multiparametric hippocampal signatures for early diagnosis of alzheimer's disease using 18F-FDG PET/MRI radiomics. CNS Neuroscience & Therapeutics. Vol. 30, pg. e14539, 2023., https://doi.org/10.1111/cns.14539.

7 A. Zwanenburg et al., The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology. Vol. 295, pg. 328–338, 2020., https://doi.org/10.1148/radiol.2020191145.

8 R. J. Gillies, P. E. Kinahan, H. Hricak, Radiomics: images are more than pictures, they are data. Radiology. Vol. 278, pg. 563–577, 2016., https://doi.org/10.1148/radiol.2015151169.

9 M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, G. Cook, Introduction to radiomics. Journal of Nuclear Medicine. Vol. 61, pg. 488–495, 2020., https://doi.org/10.2967/jnumed.118.222893.

10 H. Xia, X. Luan, Z. Bao, Q. Zhu, C. Wen, M. Wang, W. Song, A multicohort study of the hippocampal radiomics model and its associated biological changes in alzheimer's disease. Translational Psychiatry. Vol. 14, pg. 111, 2024., https://doi.org/10.1038/s41398-024-02836-9.

11 T. T. Yin, M. H. Cao, J. C. Yu, T. Y. Shi, X. H. Mao, X. Y. Wei, Z. Z. Jia, T1-weighted imaging-based hippocampal radiomics in the diagnosis of alzheimer's disease. Academic Radiology. Vol. 31, pg. 5183–5192, 2024., https://doi.org/10.1016/j.acra.2024.06.012.

12 W. Hu, Q. Du, L. Wei, D. Wang, G. Zhang, Novel multi-task learning for alzheimer's stage classification using hippocampal mri segmentation, feature fusion, and nomogram modeling. European Journal of Medical Research. Vol. 30, pg. 910, 2025., https://doi.org/10.1186/s40001-025-03204-z.

13 M. Yang, S. Meng, F. Wu, F. Shi, Y. Xia, J. Feng, J. Zhang, C. Li, Automatic detection of mild cognitive impairment based on deep learning and radiomics of mr imaging. Frontiers in Medicine. Vol. 11, pg. 1305565, 2024., https://doi.org/10.3389/fmed.2024.1305565.

14 L. Wang, Q. Feng, X. Ge, F. Chen, B. Yu, B. Chen, Z. Liao, B.-L. Lin, Y. Lv, Z. Ding, Textural features reflecting local activity of the hippocampus improve the diagnosis of alzheimer's disease and amnestic mild cognitive impairment: a radiomics study based on functional magnetic resonance imaging. Frontiers in Neuroscience. Vol. 16, pg. 970245, 2022., https://doi.org/10.3389/fnins.2022.970245.

15 Q. Feng, Q. Song, M. Wang, P. Pang, Z. Liao, H. Jiang, D. Shen, Z. Ding, Hippocampus radiomic biomarkers for the diagnosis of amnestic mild cognitive impairment: a machine learning method. Frontiers in Aging Neuroscience. Vol. 11, pg. 323, 2019., https://doi.org/10.3389/fnagi.2019.00323.

16 Q. Zheng, Y. Zhang, H. Li, X. Tong, M. Ouyang, How segmentation methods affect hippocampal radiomic feature accuracy in alzheimer's disease analysis? European Radiology. Vol. 32, pg. 6965–6976, 2022., https://doi.org/10.1007/s00330-022-09081-y.

17 R. N. Mitchell-Hay, T. S. Ahearn, A. D. Murray, G. D. Waiter, Investigation of the inter- and intrascanner reproducibility and repeatability of radiomics features in T1-weighted brain MRI. Journal of Magnetic Resonance Imaging. Vol. 56, pg. 1559–1568, 2022., https://doi.org/10.1002/jmri.28191.

18 Y. Li, S. Ammari, C. Balleyguier, N. Lassau, E. Chouzenoux, Impact of preprocessing and harmonization methods on the removal of scanner effects in brain mri radiomic features. Cancers. Vol. 13, pg. 3000, 2021., https://doi.org/10.3390/cancers13123000.

19 S. Leandrou, D. Lamnisos, H. Bougias, N. Stogiannos, E. Georgiadou, K. G. Achilleos, C. S. Pattichis, A cross-sectional study of explainable machine learning in alzheimer's disease: diagnostic classification using mr radiomic features. Frontiers in Aging Neuroscience. Vol. 15, pg. 1149871, 2023., https://doi.org/10.3389/fnagi.2023.1149871.

20 M. Velazquez, Y. Lee, Multimodal ensemble model for alzheimer's disease conversion prediction from early mild cognitive impairment subjects. Computers in Biology and Medicine. Vol. 151, pg. 106201, 2022., https://doi.org/10.1016/j.compbiomed.2022.106201.

21 C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. Vol. 1, pg. 206–215, 2019., https://doi.org/10.1038/s42256-019-0048-x.

22 S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Vol. 30, pg. 4765–4774, 2017., https://papers.nips.cc/paper_files/paper/2017.

23 Y. Du, S. Zhang, Y. Fang, Q. Qiu, L. Zhao, W. Wei, Y. Tang, X. Li, Radiomic features of the hippocampus for diagnosing early-onset and late-onset alzheimer's disease. Frontiers in Aging Neuroscience. Vol. 13, pg. 789099, 2022., https://doi.org/10.3389/fnagi.2021.789099.

24 D. Miao, X. Zhou, X. Wu, C. Chen, L. Tian, Hippocampal morphological atrophy and distinct patterns of structural covariance network in alzheimer's disease and mild cognitive impairment. Frontiers in Psychology. Vol. 13, pg. 980954, 2022., https://doi.org/10.3389/fpsyg.2022.980954.

25 C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish, E. Liu, J. L. Molinuevo, T. Montine, C. Phelps, K. P. Rankin, C. C. Rowe, P. Scheltens, E. Siemers, H. M. Snyder, R. Sperling, C. Elliott, E. Masliah, L. Ryan, N. Silverberg, NIA-AA research framework: toward a biological definition of alzheimer's disease. Alzheimer's & Dementia. Vol. 14, pg. 535–562, 2018., https://doi.org/10.1016/j.jalz.2018.02.018.

26 B. H. Brinkmann, H. Guragain, D. Kenney-Jung, J. Mandrekar, R. E. Watson, K. M. Welker, J. W. Britton, R. J. Witte, Segmentation errors and intertest reliability in automated and manually traced hippocampal volumes. Annals of Clinical and Translational Neurology. Vol. 6, pg. 1807–1814, 2019., https://doi.org/10.1002/acn3.50885.

27 M. Panahi, M. Habibi, M. S. Hosseini, Enhancing mri radiomics feature reproducibility and classification performance in parkinson's disease: a harmonization approach to gray-level discretization variability. Magnetic Resonance Materials in Physics, Biology and Medicine. Vol. 38, pg. 23–35, 2025., https://doi.org/10.1007/s10334-024-01215-1.

28 S. Zhan, J. Wang, J. Dong, X. Ji, L. Huang, Q. Zhang, D. Xu, L. Peng, X. Wang, Y. Zhang, S. Liang, L. Chen, Machine learning prediction prior to onset of mild cognitive impairment using t1-weighted magnetic resonance imaging radiomic of the hippocampus. Asian Journal of Psychiatry. Vol. 108, pg. 104532, 2025., https://doi.org/10.1016/j.ajp.2025.104532.

© The National High School Journal of Science 2026