Supplementary Material

Classifying Alzheimer's Disease Neuropathology Using Clinical and MRI Measurements

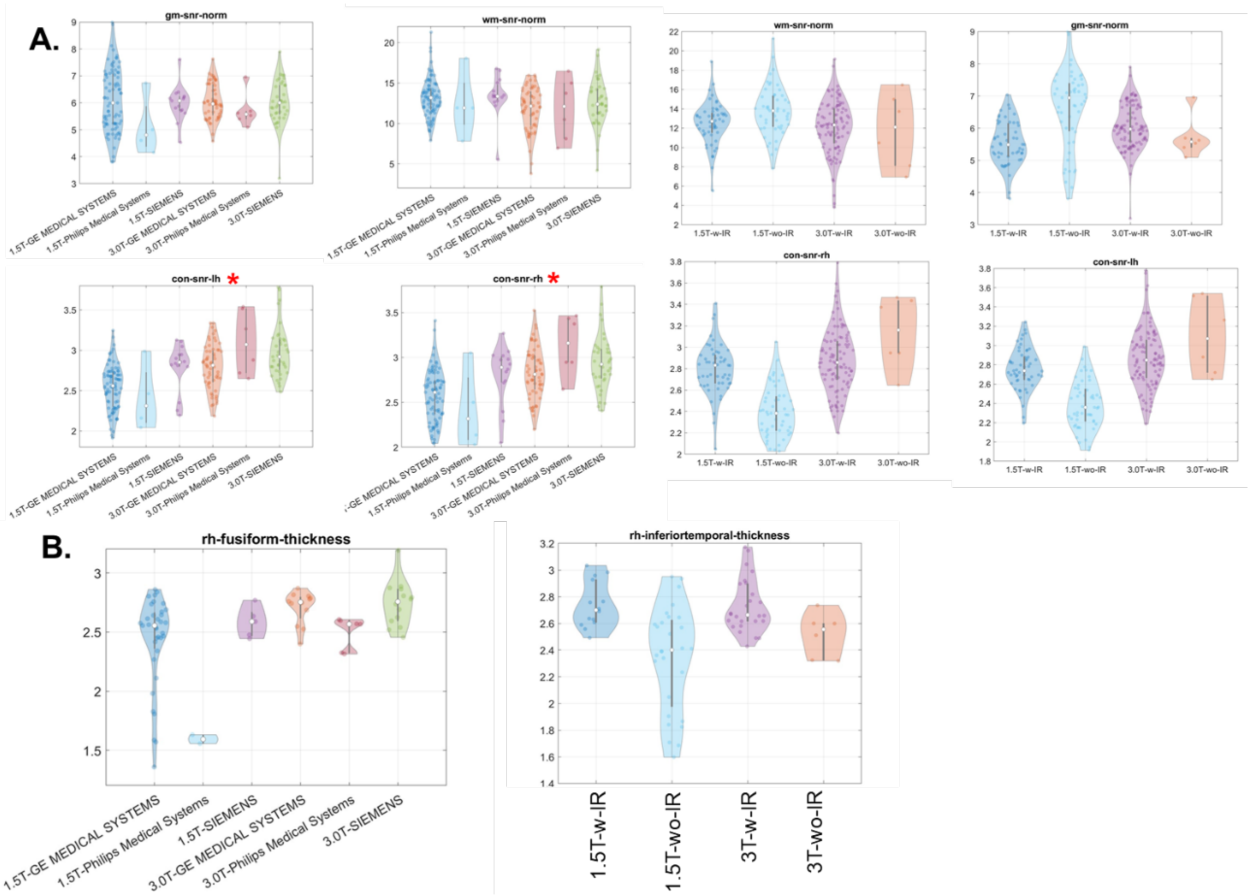
A. Data-driven method for FreeSurfer quality control

For each MRI scan, we obtained scanner field strength, scanner manufacturer, and scanning protocols from the DICOM header. We particularly obtained scan resolution and inversion-recovery (IR) time from the scanning protocols. The maximum acceptable slice thickness and in-plane resolution for these scans were 1.5 mm and 1.5 mm x 1.5 mm, respectively.

The FreeSurfer processing pipeline failed on a total of 130 participant scans for various reasons, with failed Talairach registration as most common; this may have been due to the arbitrary removal of scan information during submission of DICOM data to NACC. Consequently, we started with 65 ADNC0, 67 ADNC1, and 266 ADNC3 participants that successfully finished FreeSurfer 6.0 in our analyses.

For these FreeSurfer runs, we utilized the fsqc toolbox to perform the quality control step. This python routine outputs several matrices as quality measures for FreeSurfer-processed structural MRI data. We focused on the signal-to-noise (SNR) ratio for white matter (WM) and grey matter (GM) in norm.mgz (wm_snr_norm and gm_snr_norm), and on the WM-to-GM contrast signal-to-noise ratio (SNR) in the left and right hemisphere (con_lh_snr and con_rh_snr), as our main FreeSurfer QC matrices.

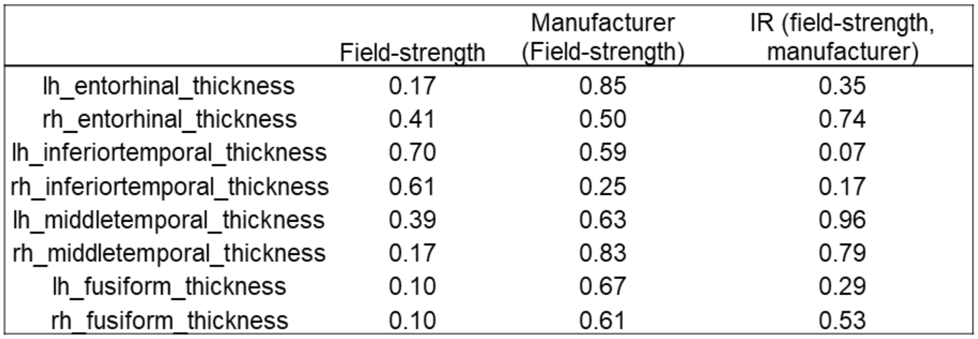
We utilized a data-driven approach to assess the data quality using QC matrices. More specifically, we conducted a repeated measures analyses of variance (rmANOVA) to determine if QC matrices were significantly different among scanner field strength, scanner manufacturer or the implementation of IR (Supplementary Figure 1A). In addition, to further evaluate whether various scanners could significantly affect FreeSurfer outputs, we conducted another rmANOVA on cortical thickness measures across scanner field strength, scanner manufacturer, and the implementation of IR in the ADNC0 group (i.e., whose structure was least affected by disease pathology (Supplementary Figure 1B). We removed any scans that produced extreme QC values and cortical thickness measures from our analyses.



Supplementary Figure 1. A) Boxplot of QC matrices across scanner field strength and manufacturer (left) and scanning protocol (implementation of IR, right) in all participants. B) Boxplot of a representative cortical thickness measures across scanner strength and manufacturer (left) and scanning protocol (implementation of IR) in ADNC0 groups.

As shown in Supplementary Figure 1, scanner field strength and manufacturer or scanning protocol (implementation of IR) did not significantly affect WM or GM SNRs, but significantly affected the WM-to-GM CNR. Therefore, in training our classification models, besides the binary feature indicating scanner field strength (1.5T (0) or 3.0T (1)), we included a categorical feature indicating scanner type (GE, Philips, or Siemens), and a binary feature indicating whether inversion recovery (IR) was utilized (1) or not (0) as features. For thickness measures, scans collected on 1.5T Philips scanner produced a significantly lower measures as compared to other scanners. In this case, we removed participants with MRI data collected on 1.5T Philips scanner from both groups in our analyses.

After we finalized our cohort, we additionally conducted a rmANOVA to evaluate if scanner type or protocol would affect the FreeSurfer thickness measures of these 8 meta-ROIs in our ADNC0&1 group. Supplementary Table 1 below lists the significance levels (p-values) for this analysis and indicates that in our finalized cohort without severe amyloid pathologies, scanner type or protocol would not significantly affect the thickness measures.



	Field-strength	Manufacturer (Field-strength)	IR (field-strength, manufacturer)
lh_entorhinal_thickness	0.17	0.85	0.35
rh_entorhinal_thickness	0.41	0.50	0.74
lh_inferiortemporal_thickness	0.70	0.59	0.07
rh_inferiortemporal_thickness	0.61	0.25	0.17
lh_middletemporal_thickness	0.39	0.63	0.96
rh_middletemporal_thickness	0.17	0.83	0.79
lh_fusiform_thickness	0.10	0.67	0.29
rh_fusiform_thickness	0.10	0.61	0.53

Supplementary Table 1. Significance levels (*p-values*) of each term in rmANOVA. A(B) indicated that factor A is nested in factor B in rmANOVA.

B. Details of replication data set

We utilized an independent, locally collected convenient sample from the Center for Neurodegeneration and Translational Neuroscience (CNTN, https://nevadacntn.org/) Center of Biomedical Research Excellence (COBRE) study as a validation data set. All CNTN participants were recruited at Cleveland Clinic Lou Ruvo Center for Brain Health Las Vegas, Nevada. The CNTN study was approved by Cleveland Clinic Institutional Review Board and all participants gave written, informed consent. The CNTN is a longitudinal, natural history study consisting of an annual clinical examination, neuropsychological assessment, MRI, and PET acquisition [1].

Participants demographics demographic, clinical, and genetic information including sex, age, years of education, race, diagnoses and *APOE* genotypes were obtained. T1-weighted MRI scans for CNTN participants were collected on a 3.0T Siemens Skyra scanner with MPRAGE sequence (TR = 2300 ms, TE = 2.96 ms, flip angle = 9°, 1mm isotropic voxel-size). The T1-weighted MRI image was input to the same FreeSurfer 6.0 pipeline that was used for the NACC participants, generating the anatomical labeling for the same 68 cortical regions and 12 subcortical ROIs. Cortical thickness measures were calculated for each cortical ROI.

Amyloid status for CNTN participants was determined using florbetapir PET (^{18}F-AV45) imaging *in vivo*. PET images were collected on a Siemens Biograph mCT PET/CT scanner after the injection of 370 MBq (±10%) of ^{18}F-AV45, at the same visit of MRI scans. The PET scans for each subject were co-registered to their T1-weighted MRI scans, and regional average standardized uptake value ratios (SUVRs) were computed in each FreeSurfer defined ROI using cerebellar reference regions. Following a previously published AV45-PET processing pipeline[2], the composite SUVR was computed by averaging the SUVRs from frontal, anterior/posterior cingulate, lateral parietal, and lateral temporal regions. Applying a cut-off value of 1.11 [2], these 144 participants were divided into an amyloid positive group ($SUVR > 1.11$, N=73) and an amyloid negative group ($SUVR \leq 1.11$, N=71). Participants' demographics are shown in Table 1B.

C. Feature importance scores in Random Forest

Gini feature importance score

Gini importance score for feature i calculates the average decrease in node impurity across all trees that include this feature.

Out-of-box (OOB) permutation-based feature importance scores

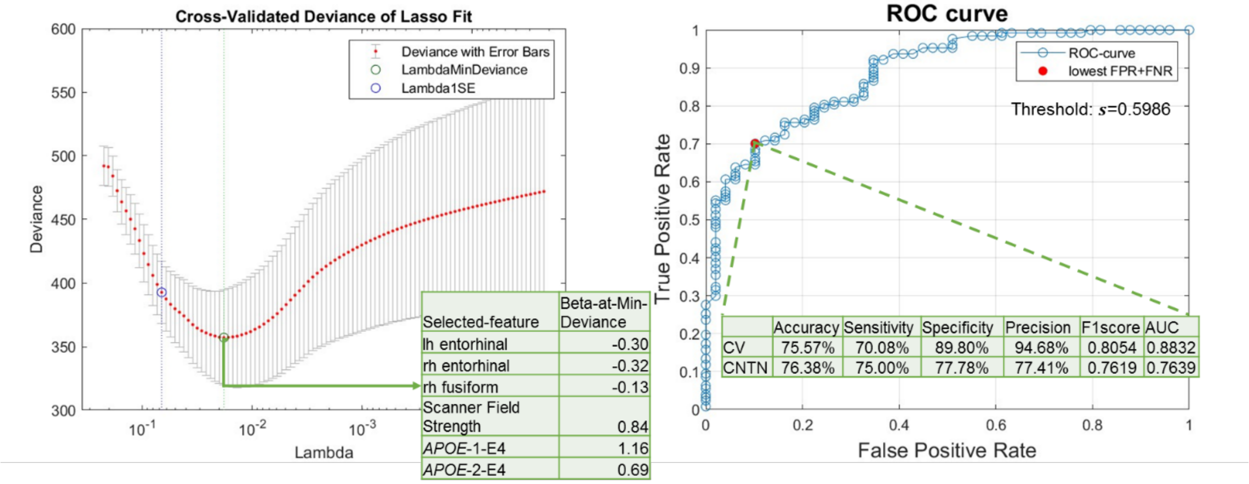
Random Forest uses the OOB samples to measure the feature importance in classification and prediction of each feature. During internal validation, OOB samples are passed down to a tree, and the classification accuracy (A_1) is recorded. Then for the same tree, values of a specific feature (i^{th} feature) are randomly permuted for OOB samples, and the corresponding accuracy (A_2) is recorded. The decrease in accuracy (A_1-A_2) is then averaged over all trees and used as a measure of OOB permutation-based importance score for this specific i^{th} feature [3]. In this case, a positive score indicating the true assignment of this feature outperforms a randomly permuted assignment, and vice versa.

D. Model performance when including ADNC3 participants with Lewy body co-pathologies

Up to 50% individuals with severe AD neuropathology (i.e., the ADNC3 group) could have some degree of Lewy bodies[4], and most of these cases with low-level Lewy bodies in the brain stem, amygdala or olfactory bulb were less likely to be presented with clinical Lewy body dementia than AD dementia. In this case, to test our model utilities with a more comprehensive real-world severe AD group, we further trained and tested our models by including participants with NPLBOD=1 (brain stem, Nsub=5), NPLBOD=4 (amygdala, Nsub=32) and NPLBOD=5 (olfactory bulb, Nsub=4) in the ADNC3 group.

With these additional ADNC3 participants, our lasso logistic regression model selected a similar set of six features as the main analyses with the minimum cross-validation error (having one or two copies of *APOE* E4 allele, cortical thicknesses encompassing fusiform and entorhinal ROIs, and scanner field strength, Supplementary Figure 2A). With these features, the cross-validation accuracy, sensitivity, specificity, precision, F1-score, and AUC were 75.57%, 70.08%, 89.80%, 94.68%, 0.81 and 0.88; and the independent testing accuracy, sensitivity, specificity, precision, F1-score, and AUC were 76.38%, 75.00%, 77.78%, 77.41%, 0.76 and 0.76 (Supplementary Figure 2B).

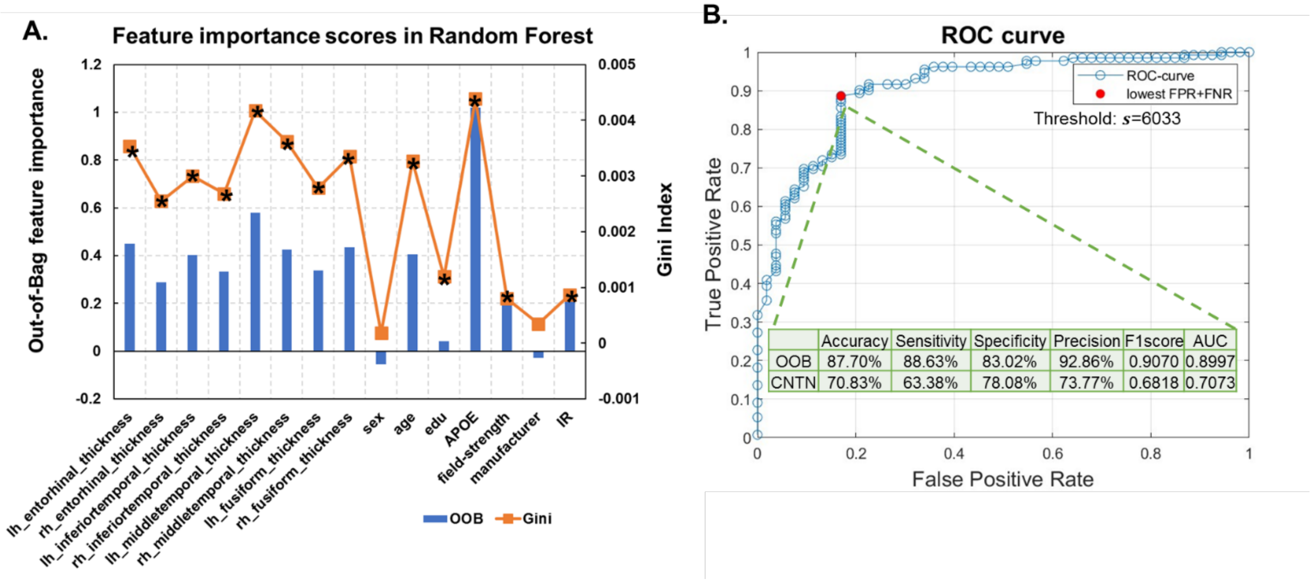
Random Forest model also gave comparable results as our main analyses, with the OOB-validation accuracy, sensitivity, specificity, precision, F1-score, and AUC being 87.70%, 88.63%, 83.02%, 92.86%, 0.91 and 0.90, respectively. The independent testing accuracy, sensitivity, specificity, precision, F1-score, and AUC were 70.83%, 63.38%, 78.08%, 73.77%, 0.69 and 0.71 (Supplementary Figure 3).



Selected-feature	Beta-at-Min-Deviance
lh entorhinal	-0.30
rh entorhinal	-0.32
rh fusiform	-0.13
Scanner Field Strength	0.84
APOE-1-E4	1.16
APOE-2-E4	0.69

	Accuracy	Sensitivity	Specificity	Precision	F1score	AUC
CV	75.57%	70.08%	89.80%	94.68%	0.8054	0.8832
CNTN	76.38%	75.00%	77.78%	77.41%	0.7619	0.7639

Supplementary Figure 2. LASSO-logistic regression results after including 41 ADNC3 participants with comorbidities of low-level Lewy bodies. A) Feature selection results. Cross-validated (CV) deviance of LASSO-logistic-regression models, trained with NACC participants to classify the presence or absence of severe AD neuropathological status (ADNC3 versus ADNC0&1), as a function of regularization strength in LASSO (lambda). The green circle corresponds to the selected model with a minimum CV deviance. The intersect table lists the beta coefficient in the logistic regression model of each selected feature. B) Model performance with selected features. ROC curve for CV performance of the reduced logistic regression model trained with 6 selected features to classify ADNC3 versus ADNC0&1. The red dot indicates the point with the lowest total false rate (false positive rate (FPR) + false negative rate (FNR)). The corresponding threshold s=0.5986 is used to binarize the predicted probability in assigning participants to the ADNC3 group. Using this model with this threshold, the intersect table shows the CV-performance with NACC participants to classify AD neuropathological status (ADNC3 versus ADNC0&1) and external testing results with CNTN participants to classify amyloid positivity status (amyloid positive versus amyloid negative).

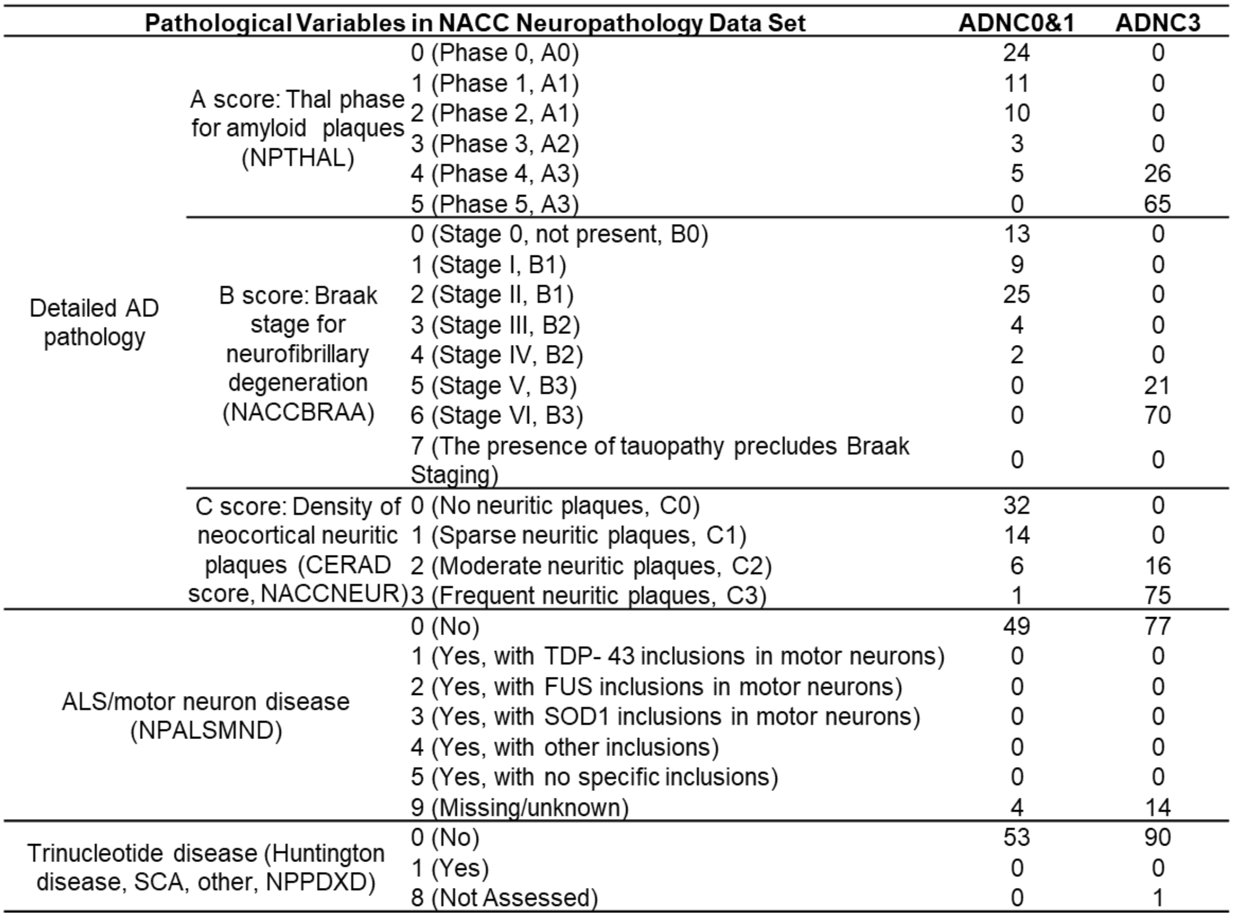


	Accuracy	Sensitivity	Specificity	Precision	F1score	AUC
OOB	87.70%	88.63%	83.02%	92.86%	0.9070	0.8997
CNTN	70.83%	63.38%	78.08%	73.77%	0.6818	0.7073

Supplementary Figure 3. Random forest results after including 41 ADNC3 participants with comorbidities of low-level Lewy bodies. A) Feature selection results. Out-of-box (OOB) permutation based (blue bars) and Gini impurity index (orange curve) based feature importance scores in the random forest model trained using all features from NACC participants to classify the presence or absence of severe AD neuropathological status (ADNC3 versus ADNC0&1). Stars (*) indicate features retained in the final model. B) Model performance with selected features. ROC curve of the random forest model with the selected features. The red dot indicates the point with the lowest total false rate (false positive rate + false negative rate). The corresponding threshold *s* =0.6033 is used to binarize the predicted probability in assigning participants to the ADNC3 group. Using this model with this threshold, the intersect table shows the OOB-validation-performance with NACC participants to classify AD neuropathological status (ADNC3 versus ADNC0&1) and external testing results with CNTN participants to classify amyloid positivity status (amyloid positive versus amyloid negative).

These results demonstrated our models' utilities when including a more comprehensive representation of real-world severe AD cases that are with comorbidities of low-level Lewy body. At the meantime, the relatively limited number of participants in the ADNC0&1 group might still preclude the emergence of significant results differences by including more ADNC3 participants. Future analyses with increased sample size could benefit from including a group of comorbidities to represent AD cases more closely in the real world.

E. Detailed AD pathology of NACC participants

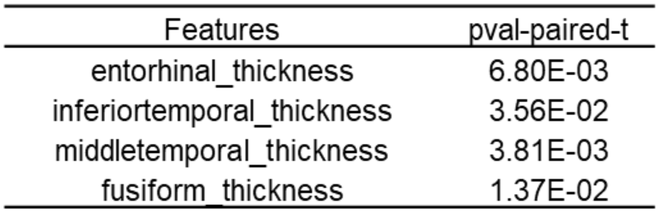


Pathological Variables in NACC Neuropathology Data Set			ADNC0&1	ADNC3
Detailed AD pathology	A score: Thal phase for amyloid plaques (NPTHAL)	0 (Phase 0, A0)	24	0
		1 (Phase 1, A1)	11	0
		2 (Phase 2, A1)	10	0
		3 (Phase 3, A2)	3	0
		4 (Phase 4, A3)	5	26
		5 (Phase 5, A3)	0	65
	B score: Braak stage for neurofibrillary degeneration (NACCBRAA)	0 (Stage 0, not present, B0)	13	0
		1 (Stage I, B1)	9	0
		2 (Stage II, B1)	25	0
		3 (Stage III, B2)	4	0
		4 (Stage IV, B2)	2	0
		5 (Stage V, B3)	0	21
		6 (Stage VI, B3)	0	70
		7 (The presence of tauopathy precludes Braak Staging)	0	0
	C score: Density of neocortical neuritic plaques (CERAD score, NACCNEUR)	0 (No neuritic plaques, C0)	32	0
		1 (Sparse neuritic plaques, C1)	14	0
		2 (Moderate neuritic plaques, C2)	6	16
		3 (Frequent neuritic plaques, C3)	1	75
ALS/motor neuron disease (NPALSMND)		0 (No)	49	77
		1 (Yes, with TDP- 43 inclusions in motor neurons)	0	0
		2 (Yes, with FUS inclusions in motor neurons)	0	0
		3 (Yes, with SOD1 inclusions in motor neurons)	0	0
		4 (Yes, with other inclusions)	0	0
		5 (Yes, with no specific inclusions)	0	0
		9 (Missing/unknown)	4	14
Trinucleotide disease (Huntington disease, SCA, other, NPPDXD)		0 (No)	53	90
		1 (Yes)	0	0
		8 (Not Assessed)	0	1

Supplementary Table 2. Detailed amyloid pathologies and comorbidities with motor disorders of NACC participants.

F. Left-right comparison between cortical thickness measures of meta-ROIs

We conducted a paired t-test to compare the thickness measures of meta-ROIs between the left and right hemisphere. For all 4 ROIs, we observed significant differences in thickness measures between the left and right hemisphere (Supplementary Table 3). Therefore, we kept both left and right hemisphere measures for each ROI in our analyses.

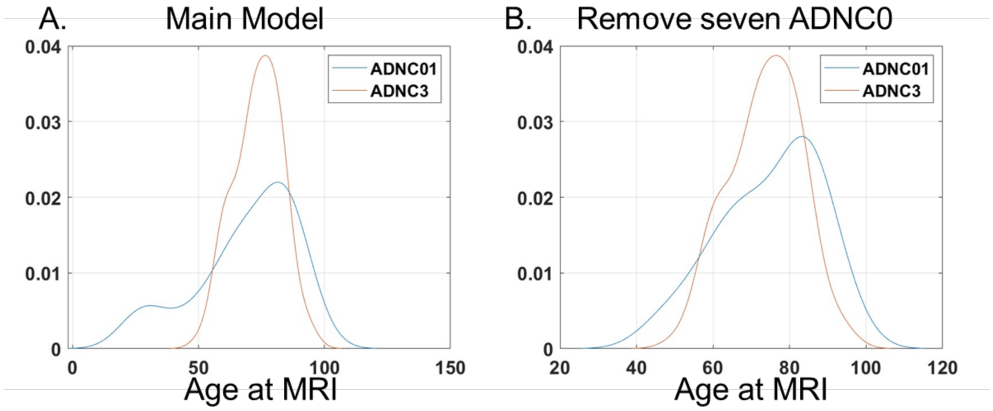


Features	pval-paired-t
entorhinal_thickness	6.80E-03
inferiortemporal_thickness	3.56E-02
middletemporal_thickness	3.81E-03
fusiform_thickness	1.37E-02

Supplementary Table 3. Significance level of the paired t-test on cortical thickness measures from 4 meta-ROIs between the left and right hemispheres.

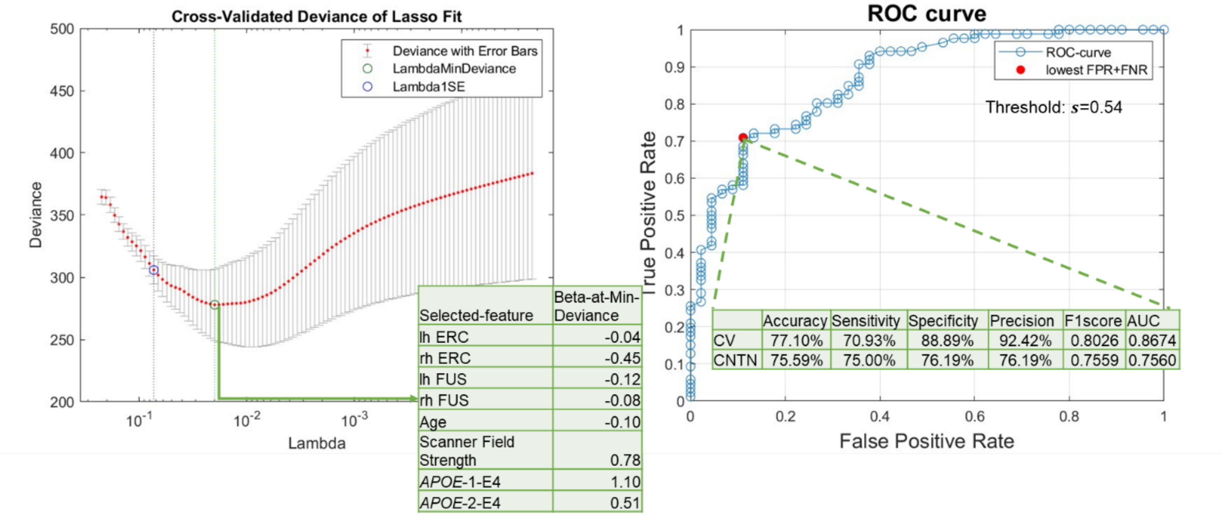
G. Repeated analyses after removing seven ADNC0 participants that were under the age of 45 at death

As shown in Table 1A, the large age variance of ADNC0&1 group (68.93±19.69), and the differences in this mean age from the overall ADNC0&1 group in NACC Neuropathology Data Set warrant further investigation. We plotted the probability density function of participant age for this group (i.e., corresponding with timing of MRI variables; (blue in Supplementary Figure 4A). This revealed seven participants under 40 years of old of at time of MRI scanning and 45 at death, respectively. Removing these seven participants resulted in a mean age of 74.89±13.00 for the ADNC0&1 group (see Supplementary Figure 4B) for the corresponding probability density map). It appears that these seven participants were responsible for the lower mean age in the ADNC0&1 group, relative to the larger neuropath cohort.



Supplementary Figure 4. Probability density maps of age at MRI variables in both groups. A) 53 and 91 ADNC0&1 and ADNC3 participants. B) After removing 7 ADNC0 participants that were below age of 40 and 45 at MRI and death, respectively.

To further evaluate if these seven participants were driving our model performances, we have repeated our analyses after removing these seven participants from the ADNC0&1 group. Supplementary Figure 5 below shows the model performance of lasso-logistic regression. Briefly, as compared to our main model (Fig. 3), similar classification accuracies were obtained on both cross-validation (77.78% versus 77.10%) and independent testing data sets (76.38% versus 75.59%). Likewise, similar findings were observed for random forest analysis. These results suggest that these seven participants did not unduly bias our main model results.



Selected-feature	Beta-at-Min-Deviance
lh ERC	-0.04
rh ERC	-0.45
lh FUS	-0.12
rh FUS	-0.08
Age	-0.10
Scanner Field Strength	0.78
APOE-1-E4	1.10
APOE-2-E4	0.51

	Accuracy	Sensitivity	Specificity	Precision	F1score	AUC
CV	77.10%	70.93%	88.89%	92.42%	0.8026	0.8674
CNTN	75.59%	75.00%	76.19%	76.19%	0.7559	0.7560

Supplementary Fig. 5. LASSO-logistic regression results after removing seven ADNC0 participants who drove the large age deviation in ADNC0&1 group. A) Feature selection results. Cross-validated (CV) deviance of LASSO-logistic-regression models, trained with NACC participants to classify amyloid pathological status (ADNC3 versus ADNC0&1), as a function of regularization strength in LASSO (lambda). The green circle corresponds to the selected model with a minimum CV deviance. The intersect table lists the beta coefficient in the logistic regression model of each selected feature. B) Model performance with selected features. ROC curve for CV performance of the reduced logistic regression model trained with 8 selected features to classify ADNC3 versus ADNC0&1. The red dot indicates the point with the lowest total false rate (false positive rate (FPR) + false negative rate (FNR)). The corresponding threshold *s*=0.54 is used to binarize the predicted probability in assigning participants to the ADNC3 group. Using this model with this threshold, the intersect table shows the CV-performance with NACC participants to classify amyloid pathological status (ADNC3 versus ADNC0&1) and external testing results with CNTN participants to classify amyloid positivity status (amyloid positive versus amyloid negative).