Received: 23 November 2019 | Revised: 15 April 2020 | Accepted: 8 June 2020

DOI: 10.1002/hbm.25115

RESEARCH ARTICLE

WILEY

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

Karteek Popuri[1] | Da Ma[1] | Lei Wang[2] | Mirza Faisal Beg[1]

[1]School of Engineering Science, Simon Fraser University, Barnaby, British Columbia, Canada

[2]Feinberg School of Medicine, Northwestern University, Evanston, Illinois

Correspondence
Mirza Faisal Beg, School of Engineering Science, Simon Fraser University, ASB 8857, 8888 University Drive, Simon Fraser University, Burnaby, BC V5A1S6, Canada.
Email: faisal-lab@sfu.ca

Funding information
Alzheimer's Disease Neuroimaging Initiative (ADNI); National Institute on Aging (R01 AG055121-01A1); Michael Smith Foundation for Health Research (MSFHR); Pacific Alzheimer's Research Foundation; Foundation Brain Canada; Canadian Institutes of Health Research (CIHR); National Science Engineering Research Council (NSERC); Alzheimer Society Research Program, Grant/Award Number: ASRP 19-09; Compute Canada

Abstract

Biomarkers for dementia of Alzheimer's type (DAT) are sought to facilitate accurate prediction of the disease onset, ideally predating the onset of cognitive deterioration. T1-weighted magnetic resonance imaging (MRI) is a commonly used neuroimaging modality for measuring brain structure in vivo, potentially providing information enabling the design of biomarkers for DAT. We propose a novel biomarker using structural MRI volume-based features to compute a similarity score for the individual's structural patterns relative to those observed in the DAT group. We employed ensemble-learning framework that combines structural features in most discriminative ROIs to create an aggregate measure of neurodegeneration in the brain. This classifier is trained on 423 stable normal control (NC) and 330 DAT subjects, where clinical diagnosis is likely to have the highest certainty. Independent validation on 8,834 unseen images from ADNI, AIBL, OASIS, and MIRIAD Alzheimer's disease (AD) databases showed promising potential to predict the development of DAT depending on the time-to-conversion (TTC). Classification performance on stable versus progressive mild cognitive impairment (MCI) groups achieved an AUC of 0.81 for TTC of 6 months and 0.73 for TTC of up to 7 years, achieving state-of-the-art results. The output score, indicating similarity to patterns seen in DAT, provides an intuitive measure of how closely the individual's brain features resemble the DAT group. This score can be used for assessing the presence of AD structural atrophy patterns in normal aging and MCI stages, as well as monitoring the progression of the individual's brain along with the disease course.

Karteek Popuri and Da Ma shared equally to the joint first authorship.

KEYWORDS

Alzheimer's disease, cross-database independent validation, dementia of Alzheimer's type, dementia score, disease progression, ensemble learning, longitudinal diagnostic stratification, magnetic resonance imaging, probabilistic classifier, prognosis prediction

1 | INTRODUCTION

Clinically diagnosed dementia of Alzheimer's type (DAT) is a neurodegenerative syndrome caused by Alzheimer's disease (AD) pathology in the brain. DAT is the most common type of dementia affecting elderly and is characterized by cognitive deterioration with atrophy typically starting in the medial temporal lobes, then progressively migrating to the rest of the brain. Reductions in memory are the first hallmark signatures, and with progressive deterioration, individuals become completely dependent on caregivers for even basic daily functions. DAT is the most common form of dementia, affecting 1 in 9 people over the age of 65 years (Alzheimer's Association, 2015) and as many as 1 in 3 people over the age of 85 (Hebert, Weuve, Scherr, & Evans, 2013). As of 2015, there was an estimated 46.8 million dementia afflicted growing to reach 131.5 million in 2050 (Prince, Comas-Herrera, Knapp, Guerchet, & Karagiannidou, 2016), projecting a very sizable effort on healthcare systems and caregivers worldwide. Biomarkers that signal the onset of pathological changes in the brain, segregate individuals based on disease severity, as well as being able to directly measure the progressive degeneration caused due to AD, are highly desirable.

Various biomarkers have been studied for AD, such as the deposition of pathological amyloid (Aβ) and tau in the cerebrospinal fluid (CSF; Fagan et al., 2007; Li, Dong, Xie, & Zhang, 2013; Maruyama et al., 2001; Ritchie et al., 2017; Tapiola et al., 2009), the brain metabolic change derived from fluorodeoxyglucose positron emission tomography (FDGPET; Lu, Popuri, Ding, Balachandar, & Beg, 2018; Popuri et al., 2018), and the structural change in the brain morphology measured from the magnetic resonance imaging (MRI; Frisoni, Fox, Jack, Scheltens, & Thompson, 2010). MRI offers a direct measurement of brain structure in exquisite detail and offers the potential to distill the MRI-visible degeneration patterns into a biomarker score that can indicate the extent to which the individual's brain is “close” to a normal control brain, or to a DAT brain. Being a commonly available modality in clinical centers worldwide, MRI-based biomarkers, once validated, offer the potential to be widely utilized.

Numerous studies have been conducted to explore the potential of MRI-based DAT biomarkers (Weiner et al., 2017). These have been facilitated by the recent availability of a large database of individuals with DAT in the Alzheimer's Disease Neuroimaging Initiative (ADNI) and other openly available databases. Initial attempts to distinguish between the extremes of normal controls (NC) and DAT have been followed by attempts to predict which of the mild-cognitive impaired (MCI) individuals will progress to DAT (Mitchell & Shiri-Feshki, 2008; Sun, van de Giessen, Lelieveldt, & Staring, 2017). A great effort was placed on developing automated methods, to scale with the demand of large numbers of images that needed to be processed, as well as harness the power of computer algorithms to detect diffuse patterns of change that may not be humanly possible to detect. Local region-of-interest (ROI) scores such as based on automatically generated medial temporal lobe atrophy using image intensity and texture (Chincarini et al., 2014) and hippocampal texture have been proposed (Sørensen et al., 2016). Coupe et al. (2015) introduced a hippocampal grading score using a patch-based framework to measure the nonlocal similarity between subject and a training population consists of healthy subjects and AD patient, and achieved an accuracy of 72.5% for 7-years’ prediction while training on ADNI dataset and test on an independent dataset. Combined hippocampal subfield volumes were shown to be more effective than the total hippocampal volume for predicting 1-year MCI converters (Khan et al., 2015). The corpus callosum's atrophy and shape change have been proposed as another identifiable feature that the converter and nonconverter MCI patient (over 1-year's period) showed significant statistical difference change rate (Elahi, Bachman, Lee, Sidtis, & Ardekani, 2015). MCI converters (over 3-year period) also showed significantly lower gray/white matter contrast compared to the nonconverter, but not for the cortical thickness and the hippocampus volume (Jefferson et al., 2015). Other studies have tried to incorporate longitudinal morphological change as features to train the classifier and improve the classification accuracy (Fiot et al., 2014; Liu, Zhou, Shen, & Yin, 2013; Lorenzi, Pennec, Frisoni, & Ayache, 2014; Sun et al., 2017; Zhang et al., 2012).

However, to provide a clinically relevant index for early diagnosis of DAT, it is important to assess a model's prediction accuracy on patient groups at various stage of the disease progression spectrum, over different ranges of time to conversion (TTC) before the disease onset. Current literatures have yet to address these two important aspects. In addition, no current studies have performed a comprehensive independent testing over populations that are assembled across all the currently available database.

Our proposed approach in this article is based on mining the whole brain 3D MRI image for patterns that measure similarity to the two extremes, the NC group on one hand that represent normal healthy aging, and the DAT group, on the other hand, have been clinically diagnosed with AD. Confidence in clinical diagnosis in these two groups is high relative to the other groups, and they represent the two extremes that are of interest; relatively lower neurodegeneration expected in the cognitively normal individuals versus the relatively fuller extent of neurodegeneration expected in DAT. Hence, these are used to train an ensemble-learning based classifier to identify these extremal patterns. The ensemble consists of a bank of classifiers assessing similarity to the NC and DAT group patterns. Each classifier generates a probabilistic score regarding membership of the given input to the two training classes. A composite ensemble score is generated across all the ensemble classifiers by averaging to represent the aggregate overall probability of belonging to the NC or the DAT groups. A low score represents "closeness" to the NC group patterns whereas a high score represents "closeness" to the DAT group patterns. This MRI-based DAT score, which we term as the MRDATS, can be interpreted as a similarity metric defined between the subject's MRI structural patterns and the DAT patterns, with low scores implying a higher similarity to nondemented structural patterns, and higher scores implying a higher similarity to demented structural patterns. Comprehensive validations of the MRDATS is performed over several independent datasets to showcase the performance of the ensemble-classifier trained on ADNI stable NC and stable DAT subjects to be able to classify unseen novel images taken from entirely separate databases.

2 | METHODS

2.1 | Experimental data

The MRI data for training and validating the proposed ensemble-learning based classifier was obtained from 5 publicly available databases, namely ADNI (Petersen et al., 2010), Australian Imaging Biomarkers and Lifestyle Study of Aging (AIBL; Ellis et al., 2009), Open Access Series of Imaging Studies (OASIS): Cross-sectional arm (OASIS1; Marcus et al., 2007), longitudinal arm (OASIS-2; Marcus, Fotenos, Csernansky, Morris, & Buckner, 2007), and Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD; Malone et al., 2013) databases. The ADNI dataset was used in the training phase, whereas images taken from the other four databases were solely used for the purpose of independent validation. Demographic details of the ADNI database are given in Tables 1 and 2 shows the demographic details of the other independent validation databases.

2.2 | Group stratification

To account for longitudinal clinical diagnoses, such as an NC individual converting to MCI, or to DAT in the future, we employed a database stratification based on the following seven subgroups (Figure 1; Popuri et al., 2018). In this stratification scheme, each *image* is assigned a membership in the form of "prefixGroup", where "Group" is the clinical diagnosis of the subject at imaging visit, and "prefix" signals past or future clinical diagnoses. For example, an image is designated as unstable NC (uNC) if the subject was assigned a NC diagnosis at that particular imaging visit, but the subject converts to MCI at a future timepoint. An image is designated as progressive NC (pNC) if the subject was assigned an NC diagnosis at that particular imaging visit, but the subject converts to DAT at a future timepoint. The early DAT (eDAT) images are associated with a current diagnosis of DAT, but this subject had received an NC or MCI status during previous visits (conversion within the study window, hence early DAT). The stable DAT (sDAT) images belong to the subjects with a consistent clinical diagnosis of DAT throughout the study window; hence, these individuals already had a confirmed diagnosis to DAT at the point of recruitment to the study.

Note that the images are associated with a clinical diagnosis. As such, one individual can span one or more clinical diagnoses, such as when converting from NC, to MCI, and to AD and therefore can have images at different clinical diagnoses. Specifically, one subject can have images that are labeled pNC → pNC → pMCI → eDAT, another can have images labeled sNC → sNC → sNC → sNC; another subject could have images labeled uNC → uNC → sMCI → sMCI, another could have images labeled pMCI → pMCI → eDAT → eDAT and still another could have images that are sDAT → sDAT → sDAT → sDAT and so on. In comparison, the traditional stratification based on subjects being identified as one of NC, MCI or DAT subgroups is sub-optimal with regards to identifying their images along an evolving clinical trajectory. To recap, we will use a novel seven-subgroup database stratification (Popuri et al., 2018) for assigning images to the following groups.

- *sNC (stable NC)*: Images belonging to the sNC group are from subjects diagnosed as NC at imaging time and stay as NC throughout the observation window of the study.
- *uNC (unstable NC)*: Images belonging to the uNC group are from subjects diagnosed as NC at imaging time and transition to MCI during the observation window of the study.
- *pNC (progressive NC)*: Images belonging to the pNC group are from subjects diagnosed as NC at imaging time and convert to DAT during the observation window of the study.
- *sMCI (stable MCI)*: Images belonging to the sMCI group are from subjects diagnosed as MCI at imaging time and remain clinically diagnosed as MCI throughout the observation window of the study.
- *pMCI (progressive MCI)*: Images belonging to the pMCI group are from subjects diagnosed as MCI at imaging time and convert to DAT during the observation window of the study.
- *eDAT (early DAT)*: Images in the eDAT group belong to subjects with a clinical diagnosis of DAT at imaging time, but had a clinical diagnosis of NC or MCI in an earlier visit, reflecting their recent conversion to DAT during the observation window of the study.
- *sDAT (stable DAT)*: Images in the sDAT group belong to subjects with a clinical diagnosis of DAT at imaging time, and DAT at earlier visits, reflecting their conversion to DAT prior to the observation window of the study.

The DAT− group of images are from those subjects that did not convert to DAT during the observational window of the study (i.e., the images in the sNC, uNC, and sMCI groups). The DAT+ group of images are from subjects that have a future clinical diagnosis of DAT (i.e., the pNC, pMCI, eDAT, and the sDAT subgroups).

By shifting the focus from subjects to images, this stratification scheme can incorporate past and future clinical diagnosis within an individual's trajectory. The task then becomes that of predicting whether an image belongs to the DAT− trajectory (that subject will not see a future diagnosis of DAT) or the DAT+ trajectory (that

TABLE 1 Stratification of ADNI images and associated demographic, clinical, and biomarker details

Dementia trajectory	Group name	Clinical diagnosis at imaging	Clinical progression	Subjects [M:F]	Images [1.5 T:3 T]	Age^{a} [years]	CSF^{a,b} [t-tau/A$\beta_{1\text{-}42}$]
DAT–^{c}	sNC:Stable NC^{d}	NC^{b}	NC → NC	197:226	929:939	76.25 (6.22)	0.38 (0.27)
DAT–	uNC:Unstable NC	NC	NC → MCI	37:23	175:59	78.12 (4.89)	0.45 (0.22)
DAT–	sMCI:Stable MCI	MCI^{b}	NC → MCI or MCI → MCI	315:220	845:1350	74.58 (7.73)	0.51 (0.42)
DAT+^{c}	pNC:Progressive NC	NC	NC → MCI → DAT	11:13	108:10	77.27 (4.23)	0.71 (0.40)
DAT+	pMCI:Progressive MCI	MCI	NC → MCI → DAT or MCI → DAT	188:133	754:274	75.43 (7.20)	0.79 (0.50)
DAT+	eDAT:Early DAT	DAT^{b}	NC → MCI → DAT or MCI → DAT	171:122	569:181	76.69 (6.87)	0.77 (0.43)
DAT+	sDAT:Stable DAT^{e}	DAT	DAT → DAT	182:148	603:372	75.70 (7.80)	0.85 (0.45)

Note: The stratification was based on two criteria, clinical diagnosis of subjects at the time of MRI image acquisition and their longitudinal clinical progression. Each image is assigned a membership of the form "prefixGroup", where "Group" is the clinical diagnosis at imaging visit, and "prefix" signals past or future clinical diagnoses. For example, an image is designated as pNC if the subject was assigned a NC diagnosis at that particular imaging visit, but the subject converts to DAT at a future timepoint. The eDAT images are associated with the diagnosis of DAT, but the subject had received NC or MCI status during previous ADNI visits (conversion within ADNI window). Whereas, the sDAT images belong to subjects with a consistent clinical diagnosis of DAT throughout the ADNI study window, hence these individuals have progressed to DAT prior to their ADNI recruitment. DAT: not on the DAT trajectory and will not get a DAT diagnosis in the ADNI window.

[a] The mean (standard deviation) age and CSF measure values within each group are given CSF measures were only available for a subset of images in each of the groups: four sNC (573), uNC (79), sMCI (634), pNC (41), pMCI (315), eDAT (154), sDAT (329).

[b] CSF, cerebrospinal fluid; DAT, dementia of Alzheimer's type; MCI, mild cognitive impairment; NC, normal controls; t-tau: total tau, A$\beta_{1\text{-}42}$: beta amyloid 1–42.

[c] DAT+: On DAT trajectory, that is, at some point in time, these subjects will be clinically diagnosed as DAT. DAT−: not on the DAT trajectory and will not get a DAT diagnosis in the ADNI window.

[d] Baseline sNC: N = 423, Age: 73.87 (5.78), MMSE: 29.06 (1.15), CSF: 0.39 (0.28). Follow-up sNC: N = 1,445, Age: 76.95 (6.17), MMSE: 29.02 (1.25), CSF: 0.37 (0.26).

[e] Baseline sDAT: N = 330, Age: 74.93 (7.83), MMSE: 23.16 (2.06), CSF: 0.84 (0.44). Follow-up sDAT: N = 645, Age: 76.08 (7.76), MMSE: 20.94 (4.61), CSF: 0.88 (0.45).

TABLE 2 Demographics summary of each stratified groups for all the independent validation dataset

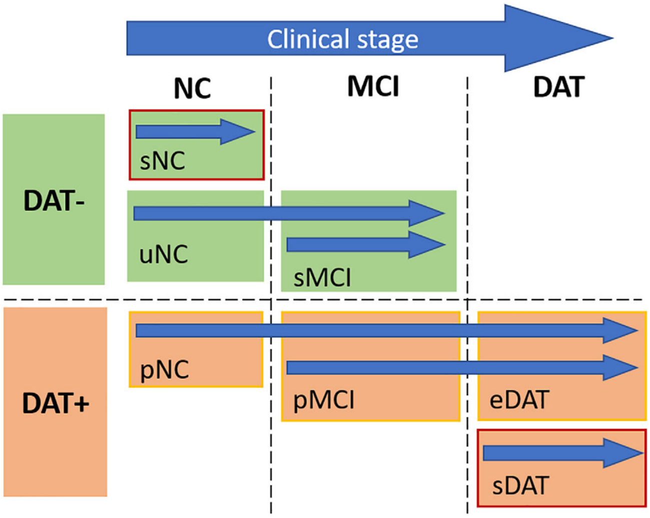
		Group name						
Dataset	Clinical measure	sNC	uNC	sMCI	pNC	pMCI	eDAT	sDAT
AIBL	Subjects [M:F] Images [1.5 T:3 T]	140:179 174:447	8:7 21:5	37:33 24:76	4:1 0:9	11:10 10:17	10:10 2:29	30:42 22:80
	Age (years)	73.45 (6.69)	72.73 (7.48)	75.97 (7.09)	73.22 (4.97)	77.78 (6.57)	79.45 (6.30)	73.79 (8.17)
OASIS-1	Subjects [M:F] Images [1.5 T:3 T]	119:197 336:0		31:39 70:0				10:20 30:0
	Age (years)	43.80 (23.75)		76.21 (7.19)				78.03 (6.91)
OASIS-2	Subjects [M:F] Images [1.5 T:3 T]	20:50 183:0	4:9 17:0	27:24 104:0		7:6 13:0	7:6 16:0	5:6 26:0
	Age (years)	76.89 (8.13)	79.34 (7.35)	78.07 (6.89)		72.69 (4.57)	74.33 (4.16)	76.31 (8.13)
MIRIAD	Subjects [M:F] Images [1.5 T:3 T]	12:11 243:0						19:27 465:0
	Age (years)	69.86 (6.94)						69.56 (6.86)

Note: Not all datasets contain all the stratified subgroups.

subject will convert to DAT in the future) regardless of the clinical diagnosis at the imaging time. The outcome of relevance is therefore not a prediction of the current clinical diagnosis associated to an image (such as NC or MCI), but whether this image is a harbinger of a future conversion to DAT (DAT+ trajectory) or if this individual will stay nondemented in the future (DAT− trajectory).

2.3 | MRI processing

The T1-weighted MRI images were segmented into the gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) tissue regions (Dale, Fischl, & Sereno, 1999) using the FreeSurfer software package version 5.3 (surfer.nmr.mgh.harvard.edu). A rigorous quality control



Longitudinal diagnosis shift	**Current** diagnosis
s: stable **u**: unstable **p**: progressive **e**: early-stage	**NC**: Normal control / cognitive normal **MCI**: Mild cognitive impairment **DAT**: Dementia of Alzheimer's Type

FIGURE 1 The schematic diagram of the group stratification. Each image is assigned a membership in the form of "prefixGroup", where "Group" is the clinical diagnosis at imaging visit, and "prefix" signals past or future clinical diagnoses. The beginning of each arrow marks the point when the participant entered the ADNI study. Each box represents the stratified group that is assigned to an *image* for the current visit of the subject. The green boxes represent the stratified groups that belong to the DAT− trajectory, while the red boxes represent the stratified groups that belong to the DAT+ trajectory. The sNC and sDAT groups are enclosed with red border indicating that their baseline images are used as the training dataset. Noted that some of the DAT− participants may switch over to DAT+ in future follow-ups

procedure was used to manually identify and correct any errors in the automated tissue segmentations following FreeSurfer's troubleshooting guidelines. Subsequently, the GM and CSF tissue regions were parcellated into 91 different anatomical ROIs using FreeSurfer's cortical (Desikan et al., 2006) and subcortical (Fischl et al., 2002) labeling pipelines. A multi-atlas segmentation approach was used to derive the total intracranial vault (TIV) segmentation for each of the MRI images (Ma et al., 2019).

2.4 | Data harmonization and W-score-based volume features

When analyzing heterogeneous data, it is important to first ensure that the data is harmonized (Fortin et al., 2017; Fortin et al., 2018; Fortin, Sweeney, Muschelli, Crainiceanu, & Shinohara, 2016; Jahanshad et al., 2013; Kochunov et al., 2015; Potvin, Dieumegarde, & Duchesne, 2017; Potvin, Mouiha, Dieumegarde, & Duchesne, 2016; Rozycki et al., 2017; Thompson et al., 2017; Yu et al., 2018), that is, removing individual variability due to such as sex, scanner field strengths, scanner type, and TIV, and only preserve differences due to effect of interest – AD-induced volume change. The data harmonization was achieved using generalized linear model (GLM) framework introduced in our previous publication (Ma, Popuri, et al., 2019) where the raw structure volume (i.e., the dependent variable) is considered as the linear combination of all other covariates (i.e., independent or predictive variables including sex, scanner field strength, scanner type, and TIV) plus a residual term (Equation (1)):

$$V_i = \beta_0 + \sum\nolimits_r^R \beta_r x_{r,i} + \varepsilon_i \qquad (1)$$

where the x_i are covariates including sex, TIV, scanner field-strength, and scanner type of each scan of subject i, and R is the total number of independent variables. The ε_i is the residual term for each image data after fitting the measurements from the reference group to the GLM using the multivariate linear regression. Here, we use the baseline measurements from the sNC group as the reference group.

The standardized residual, also known as the w-score (Equation (2)) was then calculated as the feature to train the machine-learning classifier for computing the MRI DAT score.

$$W_i = \left(\varepsilon_i - \mu_{\varepsilon_{sNC}}\right) / \sigma_{\varepsilon_{sNC}} \qquad (2)$$

where $\mu_{\varepsilon_{sNC}}$ and $\sigma_{\varepsilon_{sNC}}$ are the mean and standard deviation of the GLM-regressed residual in the reference group (sNC). The standardized residual has been shown to better identify group-dependent differences when assessing the structural changes such as atrophy (Collij et al., 2016; la Joie et al., 2012; O'Brien & Dyck, 1995).

2.5 | MRI DAT score computation via supervised ensemble learning

We adopted the well-established ensemble learning framework to compute the proposed MRI-based DAT score from the w-score volume features (Dietterich, 2000), which combines multiple trained classifiers to achieve a more robust classification instead of relying on the outcome of a single classifier. We previously applied such an ensemble learning framework to derive fluorodeoxyglucose positron emission tomography (FDG-PET) imaging-based predictive scores for early identification of the MCI to AD converters in the ADNI database, and achieved the state-of-the-art performance (Popuri et al., 2018). The baseline images from the sNC (N = 423) and sDAT (N = 330) groups in the ADNI database were used for training the classifier as those represent the extremes of the disease spectrum with a high clinical diagnosis certainty. The baseline images from the sDAT group of images are the ones in later stages of DAT and hence represent the DAT+ group. The baseline images from the sNC represent the nondemented (DAT−) group.

In the training phase, we extracted volume w-scores of the 91 FreeSurfer ROI volumes as to the training features. To prevent over-fitting to the training data, we used the sub-bagging technique (Buhlmann, Akritas, & Politis, 2003) by randomly generating $F = 100$ subsets of the training data, with a sampling ratio of 0.8. Data were sampled in a stratified fashion by choosing the same number of samples from each class (based on the smaller of the two classes in training; here sDAT) to keep the classes balanced in each training strata, that is, $N_{train} = 2 \times \lfloor 0.8 \times 330 \rfloor = 528$ samples in each of the F training subsets.

A probabilistic multi-kernel classifier, Variational Bayes Probabilistic Multi-Kernel Learning (VBpMKL; Damoulas & Girolami, 2008a) was trained on each the F subsets of the training data. In this method, each probabilistic classifier was defined as a kernel (Shawe-Taylor & Cristianini, 2004)—A function to map the data onto high-dimensional feature space to achieve linear separation of the features using the Kernel trick (Scholkopf & Smola, 2018). The VBpMKL classifier is able to apply different kernels (e.g., Gaussian, second-order polynomial) from each feature space and embed into one single composited feature space, and learns the weight for each kernel for different feature through the variational Bayesian approximation without the need for explicit parameter tuning, and output a probabilistic estimation to each class for each data.

To obtain the largest reduced dimensionality and avoid overfitting, the maximum number of $K_{\max}$ features that can be used to train the classifier with an N_{train} example in the training dataset is $K_{\max} = N_{\text{train}} \times 2p(e)$ where $p(e)$ is the probability of error (Fitzpatrick & Sonka, 2000; Loew, 2000). To keep $p(e)$ below 5%, $K_{\max} = N_{\text{train}} \times (2 \times 0.05) = N_{\text{train}}/10$ (Raamana, Weiner, Wang, & Beg, 2015). Therefore, the top $k = \lfloor N_{\text{train}}/10 \rfloor = 52$ ROIs with the most discriminative features were selected based on a ranking of the t-statistic in each of these F subsets of training data to reduce the input feature dimension from 91 to 52. The training output of each individual probabilistic kernel classifier gives the probability $p_i \in [0\ 1]$, $i = \{1,\ldots,F\}$ that the input feature belongs to the positive (DAT+) class. We then define the MRI DAT score (in short, the MRDATS) as the mean of all the probabilistic predictions over the entire F classifiers.

The classifier ensemble output is a scalar number MRDATS that can be interpreted as a similarity measure of the volumetric MRI-derived features to the sNC and the sDAT groups in the training set. In a clinical context, the continuous score can be evaluated in conjunction with other clinical variables and need not be thresholded at an arbitrary midway threshold. However, for the purposes of evaluation, the MRDATS can be thresholded at 0.5 to create a diagnostic label of DAT− or DAT+ (either the image is from DAT+ trajectory or it is not).

Once a class label of DAT+/DAT− has been assigned, sensitivity, specificity, accuracy, and balanced accuracy are obtained by comparing to the clinical diagnosis. The trained ensemble model was then evaluated on the remaining stratified subgroups, where pNC, pMCI, eDAT, and sDAT are deemed members of DAT+ trajectory whereas sNC, uNC, and sMCI are not. Balanced accuracy, an average of sensitivity and specificity, is a better reflection of accuracy of classification in the presence of class imbalance. The area under the curve is also calculated by scanning the threshold from 0 to 1 and is an indication of the separation of the class (DAT+/DAT−) histograms.

2.6 | Evaluation on the ensemble-learning-based MRDATS prediction on independent datasets

The performance of the classifier is evaluated over independent test images taken from four publicly available datasets: AIBL, OASIS-1, OASIS-2, and MIRIAD. The demographic details of these databases are given in Table 2. MRI processing and w-scoring were performed using the same procedures performed on the ADNI data. The w-scores of volumes from the same ROIs were used as features and fed to the F-ensemble classifier trained on ADNI training data and the MRDATS were obtained for each test image. Note that not only were these databases never included in training, the demographics and scanning parameters are slightly different from each, and therefore provide a good validation of the generalizability of the MRDATS to new samples.

2.7 | Comparison between MRDATS with CSF biomarker

The t-tau to beta-amyloid ratio (t-tau/Aβ_{1-42}) in the CSF is a potential biomarker of AD pathology (Fagan et al., 2007). Therefore, we evaluated the proposed MRDATS by analyzing the correlation between the MRDATS and the t-tau/Aβ_{1-42} to assess whether the proposed structural imaging-based score is correlated to a measure of pathology.

2.8 | Assess the effect of demographic- and scanner-related variables on MRDATS

It has been shown that both demographic-related (sex and age) as well as the scanner-related (field strength) variables might influence the outcomes of an analysis involving MRI-based volume features (Ma, Popuri, et al., 2019). Therefore, we explored the effect of demographic-related (sex and age) as well as the scanner-related variables (field strength) on MRDATS.

3 | RESULTS

3.1 | W-score harmonization

The w-scores of each ROI volume were obtained by regressing out the influence of scanner field strength, sex, and TIV to harmonize the databases against these variables. Figure 2 shows a visualization of the w-score volume feature for each stratified group before (top panel, raw volumes as heatmap) and after (bottom panel, w-scores as heatmap) the data harmonization, with each column being a single subject and each row being a single ROI. The figure shows that within-group variation is greatly reduced after harmonization with the w-score, and therefore, signals of structural atrophy important for intergroup discrimination are likely enhanced.

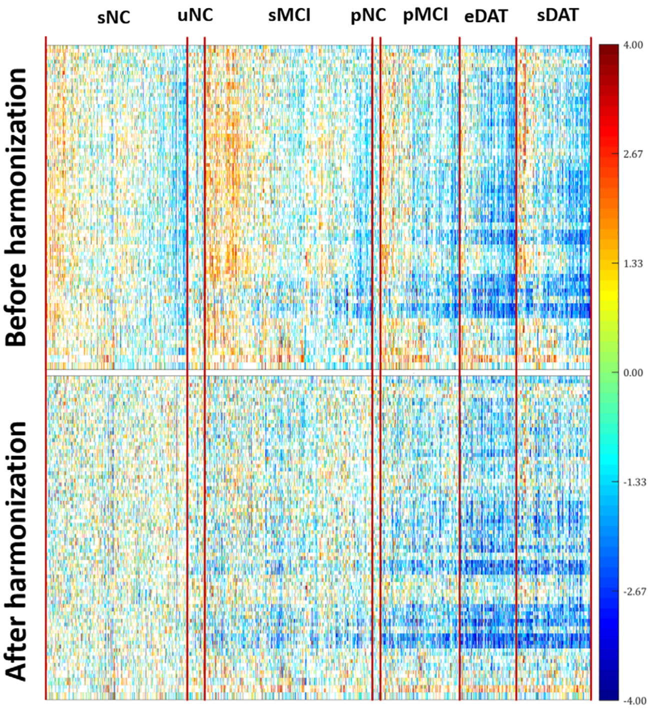


FIGURE 2 Visualization of the ROI volumes taken from the ADNI database (total of 7,168 images). *Top panel*: Raw volumes before covariate harmonization; *bottom panel*: w-score of the raw volume with respect to the sNC group, after harmonization. Each column represent one subjects' image, and each row represents one ROI in the brain

3.2 | Salient ROI selection for MRDATS computation

Automatic selection of the 52 most discriminative ROIs was achieved through a t-statistic based feature selection procedure by each of the individual classifiers in the ensemble. Table 3 lists the overall set of ROIs that were selected by the classifier ensemble along with the corresponding selection frequency for each selected ROI. The ROI selection frequency represents the percentage of times an ROI was selected by the classifiers in the ensemble model. Figure 3 shows a surface-based visualization of the selection frequency for each of the cortical ROIs.

3.3 | MRDATS distribution among training groups (sNC and sDAT)

During the training phase, only the baseline images from subjects with the most diagnostic certainty, namely, the sNC and sDAT groups, were used to train the ensemble classifier. The MRDATS values for these baseline images were determined using the out-of-bag prediction approach to avoid biased estimates (Popuri et al., 2018). In this approach, the MRDATS for a given baseline image was computed by only fusing predictions from classifiers in the ensemble that did not have the given baseline image as part of their subagging training subset. Figure 4, top panel, shows the histogram distribution of the MRDATS among the sNC and sDAT groups, that is, the classifier is used to predict the scores for the data it was trained with. The histogram shows considerable separation between the DAT+ (blue) and

TABLE 3 Most discriminative ROIs determined by the ensemble classification model

ROI Name	Frequency (%) [left | right]	ROI Name	Frequency (%) [left | right]
Accumbens-area	100.00 | 100.00	Temporal pole	100.00 | 100.00
Amygdala	100.00 | 100.00	Lateral occipital	96.00 | 100.00
Banks sts	100.00 | 100.00	Isthmus cingulate	93.00 | 95.00
Entorhinal	100.00 | 100.00	Lateral orbitofrontal	94.00 | 77.00
Fusiform	100.00 | 100.00	Insula	56.00 | 92.00
Hippocampus	100.00 | 100.00	Putamen	20.00 | 96.00
Inferior–lateral-ventricle	100.00 | 100.00	Pars orbitalis	25.00 | 87.00
Inferior parietal	100.00 | 100.00	Third-ventricle	100
Inferior temporal	100.00 | 100.00	Thalamus	31.00 | 66.00
Lateral-ventricle	100.00 | 100.00	Caudal middle frontal	80.00 | 3.00
Middle temporal	100.00 | 100.00	Posterior cingulate	66.00 | 2.00
Para hippocampal	100.00 | 100.00	Medial orbitofrontal	0.00 | 42.00
Precuneus	100.00 | 100.00	Lingual	31.00 | 2.00
Rostral middle frontal	100.00 | 100.00	Pars triangularis	21.00 | 0.00
Superior frontal	100.00 | 100.00	Postcentral	19.00 | 2.00
Superior parietal	100.00 | 100.00	Precentral	3.00 | 0.00
Superior temporal	100.00 | 100.00	Transverse temporal	0.00 | 1.00
Supramarginal	100.00 | 100.00		

Note: The ROIs are listed in descending order of their total (left and right averaged) selection frequency.

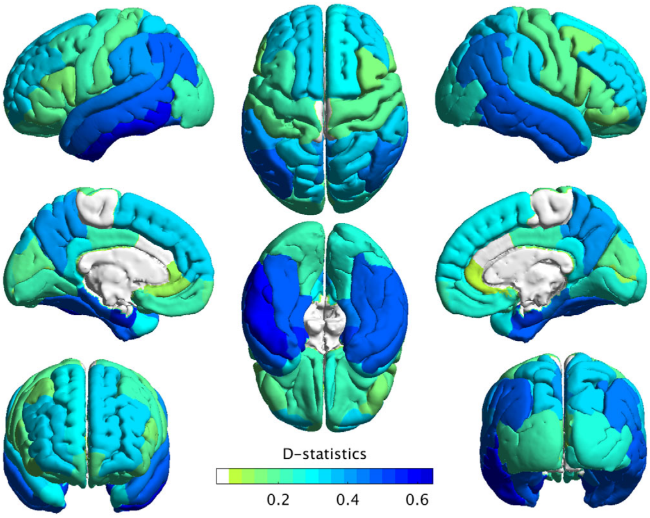


FIGURE 3 Visualization of the cortical ROIs chosen by the ensemble classification model as being the most discriminative for sNC versus sDAT. The ROIs are colored in decreasing order of their D-statistic, a measure of separation between the ECDFs of the sNC and the sDAT volume w-score measures

DAT– (green) classes, with the mean MRDATS being 0.154 for the sNC images, and 0.818 for the sDAT images. The AUC is 0.952 and approximately 60% of the sNC group was assigned a MRDATS below 0.1, and slightly less than 60% of the sDAT group was assigned an MRDATS above 0.9. The sensitivity is 0.873 specificity is 0.913 accuracy is 0.895 and balanced accuracy is 0.893.

The follow-up images of the subjects in the sNC and sDAT groups were then analyzed by the same classifier trained on the baseline sNC and sDAT images. These images are correlated with the training group of images; therefore, this is just a better estimate of the training error since the longitudinal follow-up images are, if only, slightly different from the training images. The mean MRDATS for the sNC follow-up images at 0.200 is higher than that for the baseline sNC images at 0.154, with a smaller 50% of the follow-up sNC group getting MRDATS score below 0.1. The mean MRDATS of follow-up sDAT group was 0.884 as compared to that of the baseline sDAT images at 0.818, with more than 70% of the follow-up sDAT images being assigned a MRDATS above 0.9 (Figure 4, bottom panel).

3.4 | MRDATS distribution in ADNI test groups

The trained ensemble model was evaluated on the remaining stratified subgroups (i.e., uNC, sMCI, pNC, pMCI, and eDAT) from the ADNI database. These images are unseen by the classifier as none of these subjects' images are used in classifier training. Figure 5, top panel, shows the distribution of the MRDATS among the unseen validation subgroups in ADNI and provides insights into the test (or generalization) performance of the classifier. A midway threshold of 0.5 on MRDATS was used to assign each image to either the DAT– or the DAT+ trajectory. The uNC and sMCI images are considered as belonging to the DAT– trajectory since they do not include a terminal clinical diagnosis of DAT. The pNC, pMCI, and eDAT belong to the DAT+ trajectory as they include a terminal clinical diagnosis of DAT. The MRDATS histogram shows a less pronounced class separation between DAT– and DAT+, with AUC being 0.964, as compared to 0.952 on the training data in Figure 4. The mean MRDATS for the DAT+ trajectory groups (0.867 for eDAT, 0.678 for pMCI) is much higher than that of the DAT – trajectory groups (0.350 for uNC, and 0.384 for sMCI). The exception is the pNC group, with a smaller sample of 118 images, which, although on the DAT+ trajectory, shows a group mean MRDATS of 0.329.

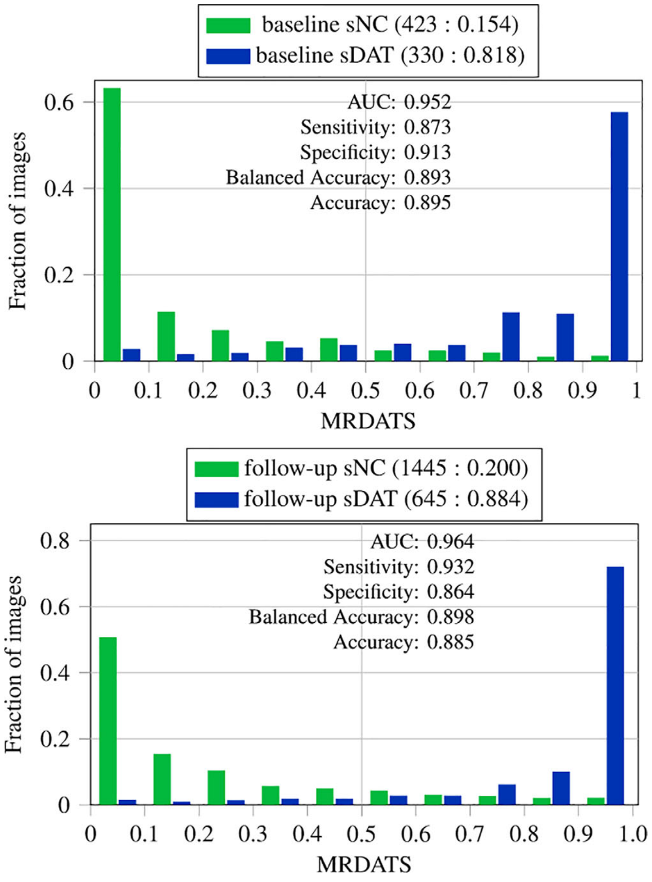


FIGURE 4 MRDATS distribution among the sNC and sDAT images and classification performance obtained in assigning images to either the DAT– or DAT+ trajectory using a binarizing 0.5 MRDATS threshold. The top panel presents the MRDATS on the baseline images used for training the ensemble model. The bottom panel shows ensemble model predictions on the follow-up images of the sNC and sDAT individuals. The follow-up images were not part of training the MRDATS computation. The (number of images: mean MRDATS) is shown for each subgroup. Balanced accuracy is the mean of the sensitivity and specificity measures

3.5 | Correlation between MRDATS and CSF t-tau/Aβ_{1-42} using ADNI data

We also investigated the association between MRDATS and CSF t-tau/Aβ_{1-42} measures through Pearson correlation analysis

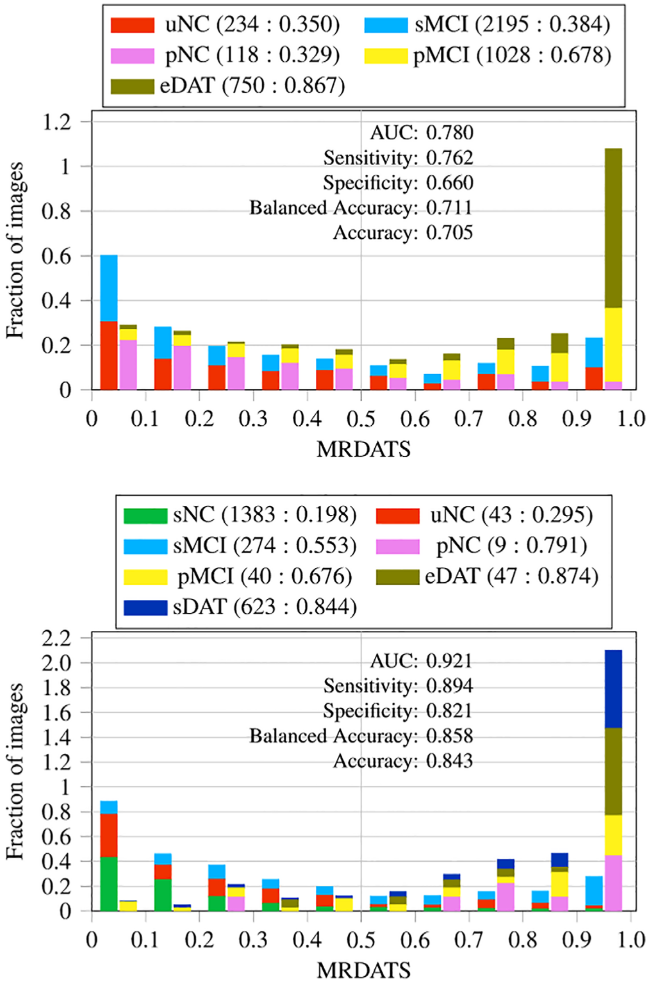


FIGURE 5 The top panel shows the MRDATS distribution among independent validation images/subjects taken from the ADNI database. The bottom panel shows the MRDATS distribution among independent validation databases namely the AIBL, OASIS-1, OASIS-2, and MIRIAD databases. The classification performance was obtained by determining dementia trajectories (DAT− or DAT+) for each image using a 0.5 MRDATS threshold. The MRDATS histograms corresponding to the DAT− (sNC, uNC, sMCI) and the DAT+ (pNC, pMCI, eDAT, sDAT) trajectories are stacked together respectively. The (number of images: mean MRDATS) for each group is shown

within the ADNI database. Figure 6 shows the correlation results for the different stratified subgroups in both the DAT+ (left column) and DAT− trajectory (right column) in the training (sNC and sDAT) and the validation (uNC, sMCI, pNC, pMCI, and eDAT) subsets of the ADNI database respectively. Significant positive correlation (p < .05) was noted between the MRDATS and the t-tau/Aβ_{1-42} for the sNC and sDAT (including both the baseline training and follow-up validation data). Similarly, a significant correlation was detected for the eDAT and sMCI subgroups. On the contrary, for more challenge cases in the uNC, pNC, pMCI, and sDAT groups, no significant correlation was detected.

3.6 | Testing using independent databases

We further evaluated the ADNI-trained ensemble models on four independent, publicly available datasets namely the AIBL, OASIS-1, OASIS-2, and MIRIAD databases. We stratify these database images also into sNC, uNC, sMCI, pNC, eDAT, and the sDAT subgroups based on their longitudinal trajectory of clinical diagnosis.

Table 2 shows the demographic information of each stratified groups for all the datasets. Not all datasets contain all the stratified subgroups. For example, OASIS-1 only contains sNC, sMCI, and sDAT, since the study is mainly focused on cross-sectional data; and MIRIAD only contains sNC and sDAT, due to the recruitment criteria as well as the short time window of the study.

The Zscape in Figure 7 showed the w-score, which is the harmonized feature of ROI volume for each stratification group, with subjects combined from all databases (ADNI, AIBL, OASIS-1, OASIS-2, and MIRIAD). The individual images (columns) in each stratification group are sorted according to their calculated MRDATS.

For each stratification group, the empirical cumulative distribution function (ECDF) of the w-score feature from the 91 ROIs volume, ranked and sorted in a descending order of the discriminative power (in terms of selection frequency as shown in Table 3), followed with the final MRDATS were shown in Figure 8. The MRDATS showed significantly improved discriminative power among different stratification groups compared to the w-scores for each individual structure.

The predicted MRDATS of each stratified group across all the test datasets is shown in Table 4. Although variation exists among different databases, there is a general trend of increasing MRDATS across the stratified groups in the order of their severity along the DAT longitudinal trajectory spectrum (severity in ascending order: sNC, uNC, sMCI, pNC, pMCI, eDAT, and sDAT) across all the databases. In addition, there is a clear distinction of the predicted MRDATS among each stratified group across all the databases, indicating potential for generalization of the MRDATS across different datasets. One exception is the pNC group. In ADNI, the mean MRDATS of pNC (0.329) is smaller than that of the sMCI group (0.384), despite the fact that it belongs to the DAT+ class versus the DAT− class that the sMCI group belongs to. On the contrary, the pNC group in AIBL showed higher MRDATS (0.791) than the pMCI group (0.677). It should be noted that the pNC group contains much smaller sample number (118 for the ADNI and 9 for the AIBL). Nevertheless, the larger variation in the MRDATS in pNC group shows the challenge of classifying this specific stratified group due to the uncertainty and variation embedded among these early-stage AD subjects.

Table 4 bottom panel shows the prediction accuracy with the proposed ensemble-learning-based classification. The stratified groups that belong to the far end of the DAT trajectory spectrum (i.e., sNC and uNC from DAT− class, and eDAT and sDAT from the DAT+ class) showed higher classification accuracy than other stratified groups that are situated in the middle of the DAT trajectory spectrum (i.e., sMCI, pNC, and pMCI), which align with the Figures 4 and 5. This is true not only for the ADNI database, in which the sNC and sDAT comprise the

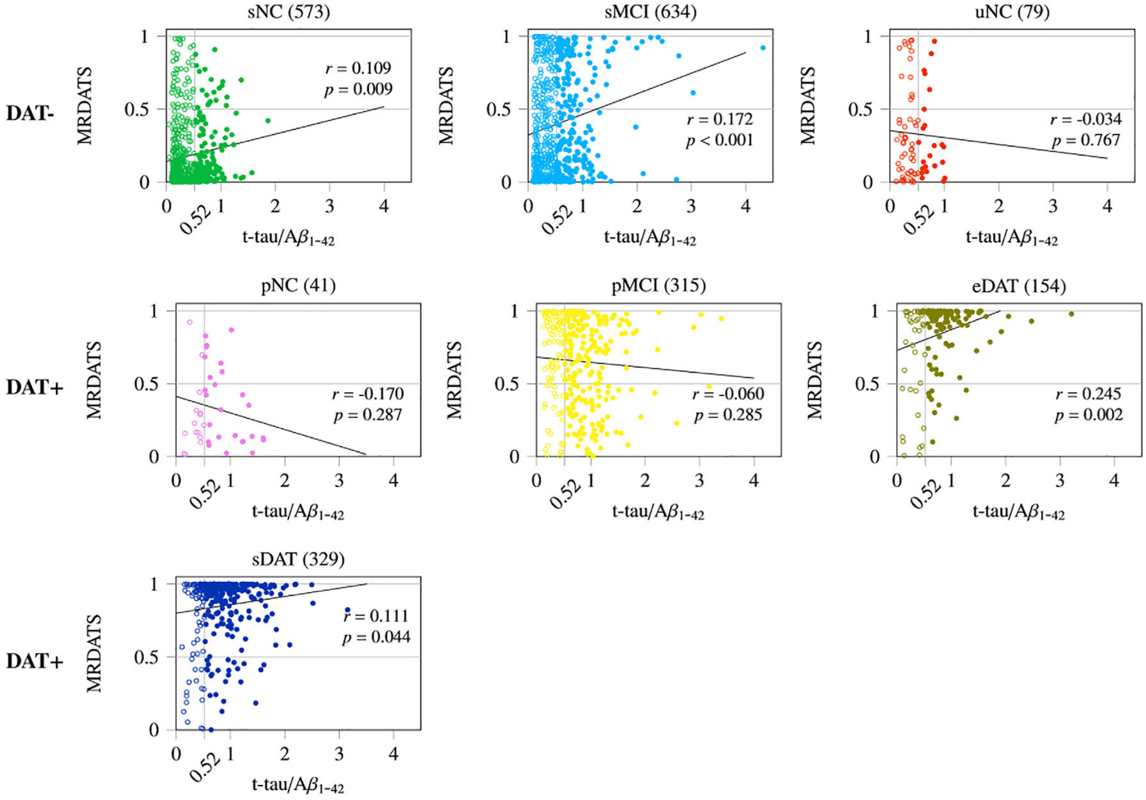


FIGURE 6 Pearson correlation between CSF t-tau/Aβ1-42 and MRDATS across different stratified groups in the ADNI database. The CSF t-tau/Aβ1-42 measures were only available for a subset of images and their numbers are shown in parentheses. A previously published 64 threshold of 0.52 was used to differentiate the low-risk (t-tau/Aβ1-42 ≤ 0.52) from the high-risk (t-tau/Aβ1-42 > 0.52) group. The solid dots represent the data with CSF t-tau/Aβ1-42 measurement equal or above the 0.52 threshold, and the hollow dots represent the CSF t-tau/Aβ1-42 measurement below the 0.52 threshold. The statistical significance threshold for correlation coefficient (*r*) was set at *p* < .05. *First row*: Correlation for the DAT− groups. *Second and third row*: Correlation for the DAT+ groups. Significant correlation between the MRI DAT score and the t-tau/Aβ1-42 were observed in the sNC, sMC group along the DAT− trajectory, and in the eDAT and sDAT in the DAT+ trajectory. Note that most of the sNC images with small t-tau/Aβ1-42 also show small MRDATS, whereas those from sDAT subgroup with high t-tau/Aβ1-42 also show high MRDATS

training data, but also for other independent validation databases, likely representing a general property of the stratified data.

An exception is again observed for the pNC group, which showed very low classification accuracy for the ADNI group (0.229), and very high classification accuracy for the AIBL group (0.889). This can also be explained both by the small sample size as well as the large variation among pNC subjects.

3.7 | MRDATS versus time to conversion in progressive image groups

Table 5 reveals the relationship between the MRDATS and the time of conversion (TTC) for the converters in the DAT+ trajectory (i.e., pNC and pMCI group), both in terms of the average MRDATS (left panel) as well as the corresponding prediction accuracy (right panel), which shows a trend towards decreasing MRDATS concomitant with a drop in prediction accuracy, as the time to conversion increases. This is likely reflecting the increasing difficulty of prognosis for images acquired earlier in the disease trajectory. In addition, comparing with the pNC group, the pMCI group consistently showed not only higher MRDATS (which matches with their definition in the DAT+ trajectory), but also higher prediction accuracy, likely because the pMCI are closer to the DAT+ side of the spectrum (an exception is the 0–1 year conversion where the pNC only has one sample subject).

3.8 | MCI conversion prediction—Comparison with state-of-the-art

Finally, we compared our methods with several state-of-the-art methods in the literature for the classification between the sMCI and pMCI subgroups within the MCI group. The AUC is used as this is the common performance metric reported among all methods. Table 6 showed a comparison of our method with some state-of-the-art methods. This comparison shows that our proposed method improves upon the state-of-the-art.

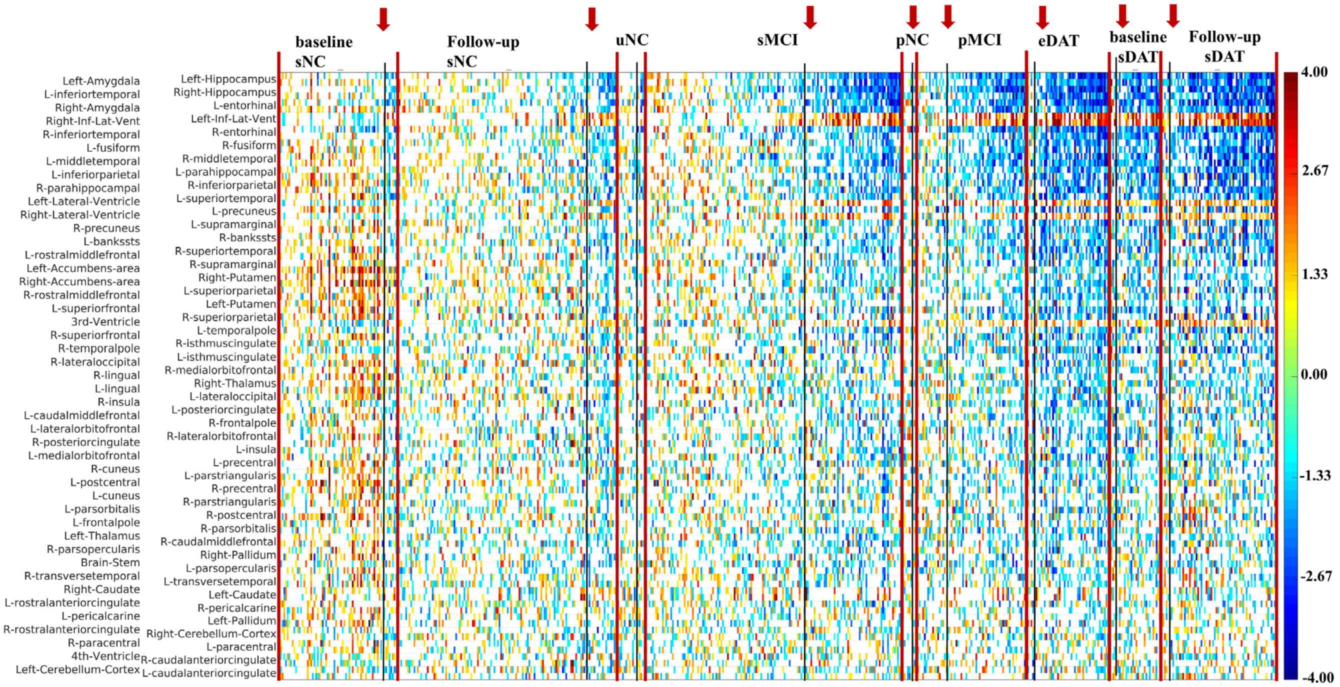


FIGURE 7 Visualization of the w-score of raw volume from 9,587 images combined across ADNI, AIBL, OASIS, and MIRIAD databases. The red arrows point to the line of 0.5 cutoff for the MRDATS for that subgroup. Each column consists of FreeSurfer-derived ROI volume-based w-scores from one subject. Each row represents one ROI in the brain across all images. Within each stratification subgroup, the images are sorted from left to right according to their MRDATS. On the vertical axis, the ROIs are sorted according to the separation of the sample distribution between the sNC and sDAT groups, calculated as the D statistic score of the K–S test. Note that the patterns for MRDATS greater than 0.5 resemble the demented (DAT+) patterns, and those less than 0.5 resemble the nondemented (DAT−) patterns. The thick red vertical line indicate the separation of different stratification groups, while the thin black vertical line with red arrow on top indicate the 0.5 cutoff point of MRDATS that separate the DAT+/DAT− patterns

4 | DISCUSSION

Brain MRI images provide a direct visualization of in vivo neuroanatomical structure that may reflect the underlying patterns of disease-specific pathology. This paper proposes a novel method to quantify the structural patterns from a brain MRI to develop a score for similarity to patterns seen in DAT images. It is important to regress out the important confounding covariates of sex (male/female), total intracranial volume, and field strength (Ma, Popuri, et al., 2019) while analyzing multicenter databases, and not analyze the raw volume features. We trained the proposed ensemble-classifier only on the subjects at the extremes of the DAT spectrum (clinically non-demented sNC vs. demented sDAT) where clinical diagnosis certainty is the highest. The trained ensemble classifier model outputs a continuous MRDATS score for a given test image, and this MRDATS score was thresholded at 0.5 to create a diagnostic label of DAT− (sNC, uNC, sMCI) or DAT+ (pNC, pMCI, eDAT, sDAT) for each of the test images. From this automated prediction, sensitivity, specificity, accuracy, and balanced accuracy measures are obtained by comparing to the clinical diagnosis.

We present a visualization of the patterns of the ROI volume w-scores across the entire MRI dataset used in this work (8,834 in total) in Figure 7. It can be observed that the baseline sNC, baseline/follow-up sDAT classes, and eDAT are found to be mostly homogeneous, with most of the subjects categorized into either the DAT− or DAT+ trajectory, respectively. The follow-up sNC and sMCI contain more proportion of DAT+ trajectory subjects compared to baseline sNC, which may indicate higher chance of the subject to develop AD at future timepoints after the current study window. Conversely, a small proportion of the pMCI subjects are associated with MRDATS that are smaller than 0.5, which may reflect the difficulty for accurate predicting the pMCI subjects. The uncertainty of diagnosis is highest in the MCI group and a larger heterogeneity in structural volume patterns exists in this group as visualized in Figure 7.

4.1 | MRDATS as a biomarker

The MRDATS is a brain MRI-based summary statistic that condenses the topographical 3D patterns within a structural T1MRI image into a single scalar value that can be interpreted as a similarity metric of the image patterns relative to those from DAT+ as compared to DAT− trajectory. Biomarkers of a progressive disease such as AD necessarily exist on a continuum. The MRDATS is a continuous scalar score, between [0–1], that mimics the continuum influence of the AD pathology in the alterations observed in brain MR structural patterns. This opens the possibility that this score can directly be used to quantify the neurodegeneration inherent in the structural patterns in the 3D MR image of the individual, and be interpreted in the context of other clinical and biomarker scores for the individual to assess their

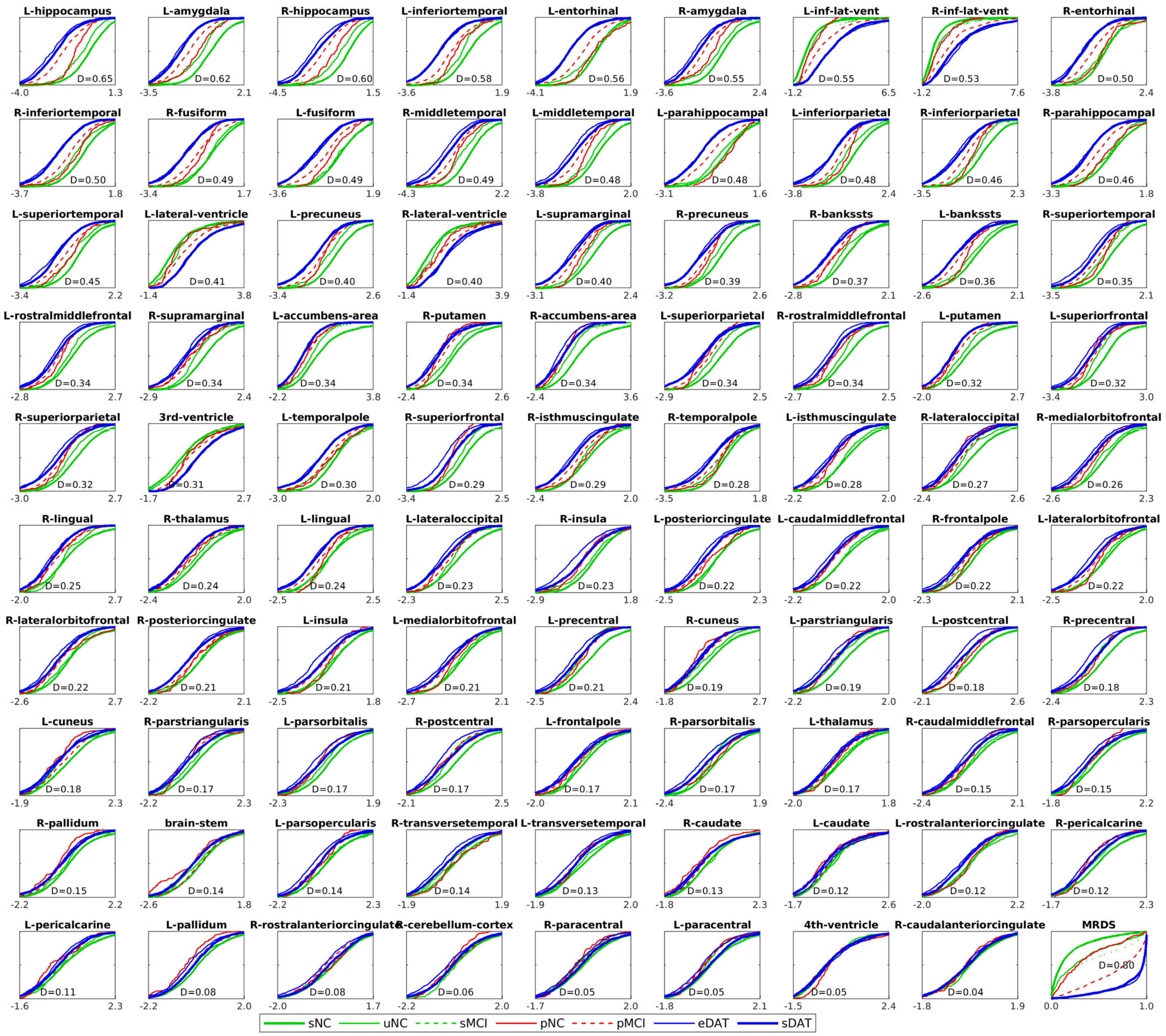


FIGURE 8 The empirical cumulative distribution function (ECDF) of the w-score feature from the 91 most discriminative ROI volumes, sorted by the separation of the sample distribution between the sNC and sDAT groups, calculated as the D statistic score of the K–S test. For each panel, the *x*-axis is the w-score in range [−4 to 4] and the *y*-axis is the ECDF of the *w*-score in the range [0, 1]. The lower right panel shows the MRDATS for each stratification group pooling all databases (ADNI, AIBL, OASIS Cross-sectional, OASIS Longitudinal, MIRIAD). Note the sDAT volume w-scores are clustered around lower values (leftward ECDFs) indicative of atrophy and reduced ROI volume relative to the sNC volume w-scores which show higher values (rightward ECDFs). This trend is reversed for the lateral ventricles which are enlarged in AD (rightward ECDFs). The sDAT MRDATS ECDFs in lower right panel are clustered toward higher values (rightward ECDFs) whereas the sNC MRDATS ECDFs are clustered towards lower values (leftward ECDFs). The separation between the ECDFs indicates the extent of separation between these stratification subgroups for that measure, and the MRDATS with a D-statistic of 0.8 shows a greater separation of the sNC/sDAT ECDFs as compared to all the raw ROI volume ECDF separations

staging along the disease spectrum. The neurodegeneration biomarker studied in Jack et al. (2017) is a surface-weighted average cortical thickness in entorhinal, inferior temporal, middle temporal, and fusiform areas. Essentially, surface-area times the average ROI cortical thickness is the volume of the ROI. Visualization of the ROI volume measures in Figure 8 shows that these are the top eight cortical ROIs (left and right) with the most separable ECDF of the w-scores of the ROI volumes between sNC and sDAT subgroups as measured by the Kolmogorov–Smirnov D-statistic. However, there are other ROIs with similar or more separable ROIs that can be used to better mark the neurodegeneration associated with AD.

A concern for biomarkers of neurodegeneration as measured via MRI for AD has been the lack of specificity; and criticism of automated image analysis algorithms has been that a small number of a priori

TABLE 4 Summary of classification performance

	ADNI	AIBL	OASIS-1	OASIS-2	MIRIAD
MRDATS					
sNC	0.200 (1445)	0.209 (621)	0.178 (336)	0.288 (183)	0.133 (243)
uNC	0.350 (234)	0.283 (26)		0.313 (17)	
sMCI	0.384 (2195)	0.478 (100)	0.617 (70)	0.581 (104)	
pNC	0.329 (118)	0.791 (9)			
pMCI	0.678 (1028)	0.677 (27)		0.673 (13)	
eDAT	0.867 (750)	0.878 (31)		0.865 (16)	
sDAT	0.884 (645)	0.831 (102)	0.839 (30)	0.848 (26)	0.847 (465)
Accuracy					
sNC	0.864 (1445)	0.878 (621)	0.938 (336)	0.792 (183)	0.951 (243)
uNC	0.714 (234)	0.846 (26)		0.765 (17)	
sMCI	0.654 (2195)	0.580 (100)	0.343 (70)	0.413 (104)	
pNC	0.229 (118)	0.889 (9)			
pMCI	0.718 (1028)	0.741 (27)		0.615 (13)	
eDAT	0.907 (750)	0.903 (31)		1.000 (16)	
sDAT	0.932 (645)	0.902 (102)	0.900 (30)	0.923 (26)	0.903 (465)

Note: Top: Predicted MRDATS of each stratified group across all the independent unseen test images in each dataset; bottom: The classification accuracy with the proposed ensemble-learning-based classification (using a 0.5 threshold). The number in the bracket shows the corresponding sample number in each stratified group.

selected ROIs are typically analyzed and therefore are insensitive to the influence of non-AD etiologies on other ROIs contributing to low specificity (Jack Jr., Barrio, & Kepe, 2013). With the availability of stable automated pipelines such as FreeSurfer for volumetric analysis, this limitation can be overcome by analyzing all the ROIs in a data-driven fashion to learn from the data which of the ROIs are most separable. The visualization in Figure 8 reveals many ROIs with separable ECDFs between sNC and sDAT, including subcortical ROIs. Hence, our ensemble classifier uses 52 of these ROIs in a purely data-driven fashion. Further, in the future, multi-class classifiers within the same framework could be trained to discriminate confounding non-AD etiologies helping overcome the nonspecificity issues.

Figure 4 shows that the ensemble-classifier-based proposed MRDATS has mostly a unimodal distribution within each sNC and sDAT subgroup, and good separation across the sNC and sDAT groups (high AUC of 0.952 and 0.964 for the baseline and follow-up images in Figure 4). There are sNC images that have higher MRDATS than 0.5 indicating that these individuals' brain MR images have a topographic 3D volumetric pattern that is more similar to the DAT+ pattern, and this is confirmed by Figure 7 that visualizes these volumetric patterns. While this could be a normal overlap of class features due to the variation across individuals in each subgroup, it may also mean that these images are indicative of preclinical AD in the brain, and may convert to DAT in the future, or these are cognitive normal individuals who carry a burden of AD pathology throughout their life with ongoing neurodegeneration but without concurrent symptoms of cognitive impairment (Aizenstein et al., 2008; Knopman, Boeve, & Petersen, 2003). Similar mirror pattern is noted for the sDAT group where most images scored higher than midway threshold but some images have lower MRDATS scores indicating more similarity to the sNC patterns, and these are also visualized in Figure 7. These observations are in concordance with reports that a range of 10% to 30% of individuals clinically diagnosed with DAT does not have AD at autopsy (Beach, Monsell, Phillips, & Kukull, 2012).

In the sMCI group, the 0.5 threshold seems to separate the sMCI group further into two distinct subsets based on their structural degeneration patterns. The sMCI with MRDATS greater than 0.5 show patterns more resembling the sDAT patterns, and these individuals may be at a higher risk of future conversion to DAT. For the pMCI, those individuals with MRDATS less than 0.5 are showing patterns more similar to the sNC patterns. They may be that select subgroup that does not have volumetric degeneration typical of the sDAT individuals. Hence, the MRDATS may be a useful construct to subdivide clinically diagnosed groups further based on their structural volumetric patterns.

Figure 5 shows the distribution of MRDATS on independent validation images taken from the ADNI database (top panel) and AIBL, OASIS, and MIRIAD databases (bottom panel). The MRDATS is lower for the subgroups on the DAT− trajectory (sNC, sMCI, uNC) and higher for the DAT+ trajectory (pMCI, eDAT, sDAT). The exception is lower MRDATS for ADNI pNC images (top panel), while higher MRDATS for the other independent validation database (bottom panel).

In this study, the MRDATS is derived from a probabilistic kernel classifier which is a generative model (Damoulas & Girolami, 2008a; Ng & Jordan, 2001). It would be interesting to also compare the performance of different models when trained using the same features, for example, comparing the generative models with the discriminative models such as Support Vector Machines (SVM) and neural networks.

TABLE 5 The effect of time to conversion (TTC) on MRDATS score for the converters in the DAT+ trajectory (i.e., pNC and pMCI groups)

	pNC	pMCI
MRDATS		
TTC (years)	0.362 (127)	0.678 (1,068)
0–1	0.537 (1)	0.770 (264)
1–2	0.479 (12)	0.712 (358)
2–3	0.472 (13)	0.669 (200)
3–4	0.440 (19)	0.593 (111)
4–5	0.386 (15)	0.545 (51)
5–6	0.358 (22)	0.573 (27)
6–7	0.285 (13)	0.395 (22)
7–8	0.306 (16)	0.477 (19)
8–9	0.213 (10)	0.398 (12)
9–10	0.123 (6)	0.293 (4)
Accuracy		
TTC (years)	0.276 (127)	0.717 (1,068)
0–1	1.000 (1)	0.826 (264)
1–2	0.417 (12)	0.757 (358)
2–3	0.385 (13)	0.730 (200)
3–4	0.368 (19)	0.586 (111)
4–5	0.267 (15)	0.549 (51)
5–6	0.273 (22)	0.556 (27)
6–7	0.231 (13)	0.364 (22)
7–8	0.250 (16)	0.474 (19)
8–9	0.000 (10)	0.417 (12)
9–10	0.000 (6)	0.250 (4)

Note: The left two columns show the MRDATS, and the right two columns show the corresponding classification accuracy. The number in the brackets shows the number of images in the particular subgroup.

4.2 | Separation and asymmetry in selected discriminative ROIs

The ECDF of the w-scores of the 91 ROI volumes are shown in Figure 8. The ROIs are arranged in descending value of the D-statistic of the K–S test, a measure of separation of the ECDFs of the sNC and sDAT groups for that ROI. One interpretation of the D-statistic could be how far along an ROI is in the disease process, over and above normal aging, as age was not a covariate in the w-score regression model. Those ROIs with a greater D-statistic were involved earlier and have been impacted more than those ROIs with a smaller D-statistic, as learned from the database. Further, there is left–right asymmetry noted in the involvement of these ROIs. In some structures, left ROIs have higher ECDF separation (higher D-statistic) whereas for others, right ROIs have higher ECDF separation. These findings are consistent with previous findings of left–right asymmetry in ROIs in AD (Shi, Liu, Zhou, Yu, & Jiang, 2009; Wachinger, Salat, Weiner, & Reuter, 2016; Yang et al., 2017).

The frequency of selection of each ROI by the classifiers in the ensemble shown in Table 3 also show a bilateral asymmetry in percentage of selection of some structures versus the others in the ensemble. For example, the selection frequency are thalamus (31% vs. 66%), putamen (20% vs. 96%), caudal middle frontal (80% vs. 3%), posterior cingulate (66% vs. 2%), medial orbitofrontal (0% vs. 42%), lingual (31% vs. 2%), and parstriangularis (21% vs. 0%). These generally agree with the separation of the ECDFs and may indicate the asymmetric influence of AD on these ROIs. Figure 3 shows a surface plot of the cortical ROIs selected by the classifier, with the D-statistic color-coding the ROIs. This visualization shows the involvement of the ROIs which is in agreement with the known evolution of AD pathology.

4.3 | Correspondence with CSF pathology

Significant correlation was found between MRDATS and CSF pathology burden via the t-tau/Aβ_{1-42}, as shown in Figure 6. Here, the cutoffs of 0.5 for MRDATS and 0.52 for t-tau/Aβ_{1-42} (Duits et al., 2014) are indicated. For the sNC and sMCI group, most individuals were both MRDATS- and t-tau/Aβ_{1-42}–. Similarly, those with MRDATS+ could be explained by concurrent t-tau/Aβ_{1-42}+ values as well. For the sDAT group, and the eDAT groups, similarly, most points cluster in quadrant with MRDATS+ and t-tau/Aβ_{1-42}+, and those with MRDATS-z can be explained by their t-tau/Aβ_{1-42}– values as well. However, no significant correlation between the MRDATS and the t-tau/Aβ_{1-42} was found in the pNC, uNC, and pMCI groups.

The CSF data reveal considerable heterogeneity in the t-tau/Aβ_{1-42} measures within the same sub-stratification groups. Importantly, subjects that are clinically diagnosed as DAT are not all t-tau/Aβ_{1-42}+. At the moment, despite the efforts that have been made to push forward the use of biological measurements as definition of Alzheimer's disease in the research framework (Jack et al., 2018), the validation of using biological measures, including CSF measurements, as biomarkers to improve the diagnosis of Alzheimer's disease is still under investigation in terms of both clinical validities (Phases 2 and 3) as well as clinical utility (Phases 4 and 5; Frisoni et al., 2017). A systematic review and meta-analysis about the CSF t-tau/Aβ_{1-42} (Ritchie et al., 2017) also noted heterogeneity in research results and concluded that there is insufficient evidence to establish the use of CSF measures for the diagnosis of AD in the current clinical practice. Specifically, the presence of abnormal CSF biomarkers in cognitively normal subject also increases with age, which pose additional challenges for its clinical validation (Toledo et al., 2015).

4.4 | Stratification within a diagnostic group

Postmortem histopathologic analyses in conjunction with clinical notes remain the gold-standard for diagnosing AD in the brain. As such, clinical diagnoses are known to be unreliable, with reports of as many as 10–30% of those clinically diagnosed with AD are found to not have AD (Beach et al., 2012), and also many individuals that have no cognitive impairment in life are postmortem found to have evidence of AD (Ossenkoppele et al., 2015). Hence, our reliance on the sNC and the sDAT images, representing the extremes of the data, is an effort to

TABLE 6 Comparison of sMCI versus pMCI classification performance obtained using MRDATS with the state-of-the-art MRI based classification methods

								Balanced	
Study	Feature type	Images [sMCI:pMCI]	Time to conversion	Evaluation scheme	AUC	Sensitivity	Specificity	Accuracy	Accuracy
Zhang et al. (2012)	Volume + PET[a,b]	50:38	0–6 months	10-fold CV	0.745	0.658	0.78	0.727	–
Zhang et al. (2012)	Volume + PET[a,c]	50:38	0–6 months	10-fold CV	0.768	0.79	0.789	0.784	–
MRDATS	Volume	2,469:77	0–6 months	INDEP.VAL	0.803	0.805	0.62	0.712	0.625
Sørensen et al. (2016)	Volume[d]	192:41	0–1 year	10-fold CV	0.705	–	–	–	–
Sørensen et al. (2016)	Texture[d]	192:41	0–1 year	10-fold CV	0.74	–	–	–	–
Sørensen et al. (2016)	Volume + Texture[d]	192:41	0–1 year	10-fold CV	0.739	–	–	–	–
Khan et al. (2015)	Volume (hippocampal subfields)[d]	90:357	0–1 year	7-fold CV	–	0.81	0.48	0.55	–
MRDATS	Volume	2,469:272	0–1 year	INDEP.VAL	0.788	0.824	0.632	0.728	0.651
Suk et al. (2017)	Volume[a]	226:167	0–1.5 years	10-fold CV	0.754	0.709	0.788	0.749	0.748
Zhu et al. (2015)	Volume[a]	56:43	0–1.5 years	5-fold CV	0.814	0.48	0.928	0.704	0.718
MRDATS	Volume	2,469:482	0–1.5 years	INDEP.VAL	0.769	0.788	0.62	0.704	0.647
Sørensen et al. (2016)	Volume[d]	140:93	0–2 years	10-fold CV	0.672	–	–	–	–
Sørensen et al. (2016)	Texture[d]	140:93	0–2 years	10-fold CV	0.742	–	–	–	–
Sørensen et al. (2016)	Volume + Texture[d]	140:93	0–2 years	10-fold CV	0.72	–	–	–	–
Chincarini et al. (2014)	Intensity + Texture[d]	103:38	0–2 years	INDEP.VAL	0.74	0.76	0.68	0.72	–
Chincarini et al. (2014)	Intensity + Texture[a]	103:38	0–2 years	INDEP.VAL	0.76	0.75	0.66	0.71	–
Cheng et al. (2015)	Volume[a]	56:43	0–2 years	10-fold CV	0.764	0.734	0.721	0.728	0.734
MRDATS	Volume	2,469:627	0–2 years	INDEP.VAL	0.76	0.783	0.62	0.701	0.653
Young et al. (2013)	Volume[d]	96/47	0–3 years	INDEP.VAL	0.643	0.532	0.698	0.615	0.643
Huang et al. (2017)	Longitudinal intensity[a]	61:70	0–3 years	10-fold CV	0.812	0.865	0.782	0.824	0.794
Liu et al. (2011, 2013)	Texture[d]	239:38	0–3 years	INDEP.VAL	0.776	0.421	0.824	0.623	0.769
Lu et al. (2018)	Volume[d]	753:409	0–3 years	INDEP.VAL	–	0.7327	0.7618	0.7473	0.7544
MRDATS	Volume	2,469:826	0–3 years	INDEP.VAL	0.751	0.777	0.62	0.698	0.659
Coupe et al. (2015)	Hippocampal grade[d,e]	309:37	0–7 years	LOOCV	0.73	0.649	0.735	0.692	–
MRDATS	Volume	2,469:1033	0–7 years	INDEP.VAL	0.729	0.733	0.62	0.676	0.653

Abbreviations: CV, cross-validation; INDEP.VAL, independent validation; LOOCV, leave-one-out cross-validation.

[a] Trained with sMCI/pMCI.

[b] Average normalized FDG-PET intensities.

[c] Longitudinal features from multiple time points.

[d] Trained with NC/AD.

[e] The Hippocampal grade is calculated based on the image similarity between the test image and the template image.

mitigate the uncertainty associated with the clinical diagnoses. The Zscapes visualization of w-scores of ROI volumes reveals that a 0.5 threshold of the proposed MRDATS can substratify each group according to the presence of dementia-like patterns. For those in the sNC group, those images with MRDATS greater than 0.5 could be showing neurodegeneration patterns without associated cognitive complaints in that individual. For those in the sDAT group, those images with MRDATS less than 0.5 are indicating relative lack of neurodegeneration patterns associated with DAT. What is considered an error on the part of the classifier for this study may be a potential limitation of the current clinical diagnoses. An MRDATS based directly in neuroanatomical volumetric patterns may be further validated to potentially enrich the groups with those that have confirmed patterns of neurodegeneration corresponding to DAT+ and those whose patterns resemble the DAT− patterns. Based on the MRDATS an intermediate transition zone corresponding to MRDATS in the range [0.4–0.6] could also be defined.

4.5 | Data imbalance

When training a machine-learning model, it is important to ensure the training data is balanced. In this work, we have ensured that the training groups are balanced to mitigate the possibility of the classifier learning a biased model. On the other hand, there is no impact of imbalance in the unseen testing groups on the performance of the classifier. Unlike most published classification studies that typically report aggregate accuracy measures, we have reported the performance measures for each of the five stratification subgroups which provide more detailed information into the classification performance of the classifier (Figure 4).

4.6 | Brain structural volume as the features for measuring DAT severity, and potential extension to other features

AD is widely considered as abnormal loss of gray matter, and brain structural volumes is strongly associated with disease severity (Frisoni et al., 2010; Jack et al., 1997; Ledig, Schuh, Guerrero, Heckemann, & Rueckert, 2018; Ma et al., 2019; Min et al., 2017; Poulin, Dautoff, Morris, Barrett, & Dickerson, 2011; Raji, Lopez, Kuller, Carmichael, & Becker, 2009; Rusinek et al., 1991; Schmitter et al., 2015; Silbert et al., 2003; Stout, Jernigan, Archibald, & Salmon, 1996; Thompson et al., 2003; Wang et al., 2019; Yi et al., 2016). In this study, we used brain structural volume as the predictive feature for measuring DAT severity to train the kernel-based classifier to predict dementia severity. Besides structural volumes, researchers have been trying to extract other types of features from MRI data to enhance the predictive power for AD, such as the image intensity and texture (Chincarini et al., 2014; Sørensen et al., 2016), cortical thinning (Bhagwat, Viviano, Voineskos, & Chakravarty, 2018; Corlier et al., 2018; Eskildsen et al., 2013; Thompson et al., 2004), or image similarity to templates (Coupe et al., 2015). These, and other features, separately, or taken together, can form natural extensions for the application of the MRDATS under the same ensemble learning framework to distill other aspects of structural information available in T1-MR images.

Specifically, brain volume and cortical thickness are two types of strongly correlated morphological features extracted from structural MRI for measuring DAT severity. A large-scale survey ((Bhagwat, Viviano, Voineskos, & Chakravarty, 2018)[1] comparing the cortical thickness and volume methods for measuring Alzheimer's disease severity showed that volume- and thickness-based measures generally perform similarly for separating clinically normal from AD populations", and volume-based measures are generally more reliable than thickness measures, although cortical thickness measurement is less correlated to, and therefore affected by, the variation of TIV. This indicates that the accuracy of TIV estimation is crucial for the effectiveness of using brain volume to measure the severity of DAT (Schwarz et al., 2016). In this study, we estimated the TIV using the multi-atlas-label-fusion method which has shown state-of-the-art results outperforming both FreeSurfer and SPM package in terms of either longitudinal consistency or test–retest reliability (Ma, Popuri, et al., 2019). The accurate estimation of TIV, along with GLM-based data harmonization methods and the use of standard residual, ensure the conversion of the structural brain MRI neurodegeneration pattern into a dementia score with strong predictive power. Nevertheless, composite features combining cortical thickness may utilize the complementary information in brain structural morphological change (Liu et al., 2011), and worth further investigated in future studies.

4.7 | Choice of feature selection methods

Feature selection can be performed in a number of ways, ranging from methods such as *t* tests (Chu, Hsu, Chou, Bandettini, & Lin, 2012; Huffman, Sobral, & Teran-Hinojosa, 2019; Wang, 2012; Wang, Zhang, Liu, Lv, & Wang, 2014; Zhou & Wang, 2007) to more complex methods utilizing PCA, mutual information, L1 norms, regression [such as LASSO (Zhao & Yu, 2006) and AdaBoost (Wang, 2012)], and so on. Our current choice of using the *t* test as a way to rank and select the most separable features is a standard method in the field as evidenced by existing literature (Damoulas & Girolami, 2008b; Shawe-Taylor & Cristianini, 2004; Varol, Gaonkar, Erus, Schultz, & Davatzikos, 2012). Evidence that this feature selection method is potentially robust to assumption violations is provided by analyzing its performance on scoring and classification of unseen test images (using top features selected during training). For these significantly large numbers of unseen test images (a total of 8,834), the calculated atrophy scores are an accurate representation of the neurodegeneration patterns (as shown by the W-score visualization in Figure 7 and ECDF in Figure 8), and images are classified demonstrated results with state-of-the-art performance (Table 6).

4.8 | Comprehensive independent testing and comparison

A comparison of our proposed MRDATS based approach with competing methods for the sMCI versus pMCI discrimination task is shown in Table 6. In the study by Coupe et al. (2015), the hippocampal grading score is calculated through the estimation of nonlocal similarity between the subject and the training data. Chincarini et al. used image similarity and texture within regions of temporal lobe atrophy as the discriminating feature and built a classifier based on PCA and SVM on the independent validation set DESCRIPA (Chincarini et al., 2014; Li et al., 2013). Khan et al. constructed classifier by combining atrophy information from multiple different hippocampal subfield, and achieve 81.1% correct prediction of pMCI but only 48.7% correct prediction of sMCI using sevenfold cross-validation (although no AUC is reported for the sMCI/pMCI classification; Khan et al., 2015). Elahi et al. (2015) used corpus callosum atrophy and shape change rate as a feature.

The strength of our validation is that we used all the images taken from the publicly available ADNI, AIBL, OASIS-1, OASIS-2, and MIRIAD databases, with a total of 8,834 images, excluding images from sNC (423) and sDAT (330) groups in ADNI baseline which are used as training data. Every image from the other databases was an independent unseen test image, and thereby the results are likely a good estimate of generalization of the classifier performance.

Despite the amazing effort from the providers of these rich databases, the typical approach from those developing novel methods has been to validate them on a subset of images to present the performance of the proposed new methods. One of the drawbacks of a limited sample validation is that it is a missed opportunity to not test any new method on the full spectrum of variations captured in these databases. Another reason to choose a smaller subset is convenience; some images may have sub-optimal initial automated segmentations. Our experience is that about 10% of the FreeSurfer automated segmentations have some variety of errors that require dedicated quality control. Our in-house anatomical team invested significant time and effort to manually correct and certify the images to be free of segmentation errors. These hurdles are representative of the real-world challenges any computational biomarker or classification algorithm will likely face when deployed in a real-world setting. Hence, our approach, to comprehensively include *all* the images from the available databases, is an effort to raise the threshold of biomarker performance reporting and thereby advance the selection of promising imaging biomarkers.

4.9 | Limitation of the current study

We note that, among all the stratified groups in the independent testing set, the MRDATS of the pNC group did not always follow the DAT trajectory. Although the pNC group is expected to be on the DAT+ trajectory, the mean predicted MRDATS for pNC subjects in the ADNI dataset (0.329) are found to be on the DAT− trajectory (<0.5), resulting in a poor predictive accuracy (0.229). One possible reason may be the smaller sample number (118) in the pNC group. Furthermore, when investigating the MRDATS as a function of time-to-conversion (TTC), a steady pattern can be observed (Table 5) showing that shorter time to conversion corresponds to higher MRDATS and higher accuracy, and both are reduced when the TTC increases. This result may indicate that the MRDATS by itself may not be enough to extract the neurodegenative patterns for the pNC group, especially when the subjects are still in the stage with longer time to conversion, and additional features such as genetic factors or CSF biomarkers might be helpful for more accurate prediction of the DAT trajectory for pNC subjects, especially during the early disease stage.

4.10 | Translation to end-users: Cloud-based validation

We have made our method available for testing on our Cloud Engine Resource for Accelerated Medical Image Computing for Clinical Applications (CERAMICCA) web portal https://ceramicca.ensc.sfu.ca. The users can upload their brain MRI scans directly on this web platform, and with a simple web-based form, launch the processing of their database. Job progress can be viewed on the website, and job control features are provided, such as viewing intermediate segmentations and/or automatically re-launching failed jobs. By hiding the complexity associated with accessing high-performance computing environments through a web-interface, the algorithms can be more easily interrogated for validation.

ACKNOWLEDGMENTS

Funding for this research is gratefully acknowledged from Alzheimer Society Research Program (ASRP 19-09), National Science Engineering Research Council (NSERC), Canadian Institutes of Health Research (CIHR), Foundation Brain Canada, Pacific Alzheimer's Research Foundation, the Michael Smith Foundation for Health Research (MSFHR), and the National Institute on Aging (R01 AG055121-01A1). We thank Compute Canada for the computational infrastructure provided for the data processing in this study.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol–Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. HoffmannLa Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.

org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data were provided in part by OASIS: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382, and OASIS: Longitudinal: Principal Investigators: D. Marcus, R, Buckner, J. Csernansky, J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, and U24 RR021382. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society. The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline and funding from the UK Alzheimer's Society (to Dr Schott) and the Medical Research Council (to Professor Fox). Professor Ourselin receives funding from the EPSRC (EP/H046410/1) and the Comprehensive Biomedical Research Centre (CBRC) Strategic Investment Award (Ref. 168). Dr Ridgway is supported by the Medical Research Council [grant number MR/J014257/1]. This work was supported by the National Institute for Health Research (NIHR) Biomedical Research Unit in Dementia based at University College London Hospitals (UCLH), University College London (UCL). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The Dementia Research Centre is an Alzheimer's Research UK (ARUK) Coordinating Centre. The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust [grant number 091593/Z/10/Z]. We are grateful to the MIRIAD participants and funders for their willingness to share these data.

CONFLICT OF INTEREST

There is no conflict of interest to declare from all authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are derived from the following resources available in the public domain: ADNI and AIBL (adni.loni.usc.edu), OASIS-1 and OASIS2 (oasis-brains.org), and MIRIAD (miriad.drc.ion.ucl.ac.uk). We have made the methods in this article available for all researchers to use and validate via our website: https://ceramicca.ensc.sfu.ca which is an online portal that connects the method with data and computes from an online form submission. We are providing access to this service to anyone at any time. Data used in preparation of this article were partly obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data used in the preparation of this article were partly obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at https://aibl.csiro.au/.

ORCID

Karteek Popuri https://orcid.org/0000-0002-1729-3255
Da Ma https://orcid.org/0000-0002-3542-7798
Lei Wang https://orcid.org/0000-0003-3870-3388
Mirza Faisal Beg https://orcid.org/0000-0003-4229-9613

ENDNOTE

[1] This study aims at developing clinical criteria and screening guidelines for AD in the pre-dementia stage. Recruiting centers were selected from EADC members in 11 European countries and included 20 memory clinics specialized in the diagnosis and treatment of memory disorders (Visser et al., 2008).

REFERENCES

Aizenstein, H. J., Nebes, R. D., Saxton, J. A., Price, J. C., Mathis, C. A., Tsopelas, N. D., … Klunk, W. E. (2008). Frequent amyloid deposition without significant cognitive impairment among the elderly. *Archives of Neurology*, *65*(11), 1509–1517. https://doi.org/10.1001/archneur.65.11.1509

Alzheimer's Association. (2015). Alzheimer's disease facts and figures. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *11*(3), 332.

Beach, T. G., Monsell, S. E., Phillips, L. E., & Kukull, W. (2012). Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer disease centers, 2005-2010. *Journal of Neuropathology and Experimental Neurology*, *71*(4), 266–273.

Bhagwat, N., Viviano, J. D., Voineskos, A. N., & Chakravarty, M. M. (2018). Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Computational Biology*, *14*(9), e1006376. https://doi.org/10.1371/journal.pcbi.1006376

Buhlmann, P. (2003). Bagging, subagging and bragging for improving some prediction algorithms. In M. G. Akritas & D. N. Politis (Eds.), *Recent advances and trends in nonparametric statistics* (pp. 9–34). Amsterdam: Elsevier.

Cheng, B., Liu, M., Suk, H.-I., Shen, D., Zhang, D., & Alzheimer's Disease Neuroimaging Initiative. (2015). Multimodal manifold regularized transfer learning for MCI conversion prediction. *Brain Imaging and Behavior*, *9*(4), 913–926. https://doi.org/10.1007/s11682-0159356-x

Chincarini, A., Bosco, P., Gemme, G., Esposito, M., Rei, L., Squarcia, S., … Nobili, F. (2014). Automatic temporal lobe atrophy assessment in prodromal AD: Data from the DESCRIPA study. *Alzheimer's & Dementia*, *10*(4), 456–467. https://doi.org/10.1016/j.jalz.2013.05.1774

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., & Lin, C. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, *60*(1), 59–70. https://doi.org/10.1016/j.neuroimage.2011.11.066

Collij, L. E., Heeman, F., Kuijer, J. P. A., Ossenkoppele, R., Benedictus, M. R., Moller, C., … Wink, A. M. (2016). Application of machine learning to arterial spin labeling in mild cognitive impairment and Alzheimer disease. *Radiology*, *281*(3), 865–875. https://doi.org/10.1148/radiol.2016152703

Corlier, F., Moyer, D., Braskie, M. N., Thompson, P. M., Dorothee, G., Potier, M.-C., … Lagarde, J. (2018). Automatic classification of cortical thickness patterns in Alzheimer's disease patients using the Louvain modularity clustering method. In E. Romero, N. Lepore, & J. Brieva (Eds.), *14th International Symposium on Medical Information Processing* *and Analysis* (Vol. 10975, p. 30). Bellingham, WA: SPIE. https://doi.org/10.1117/12.2511573

Coupe, P., Fonov, V. S., Bernard, C., Zandifar, A., Eskildsen, S. F., Helmer, C., … Collins, D. L. (2015). Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis. *Human Brain Mapping*, *36*(12), 4758–4770. https://doi.org/10.1002/hbm.22926

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194.

Damoulas, T., & Girolami, M. A. (2008a). Probabilistic multi-class multi-kernel learning: On protein fold recognition and remote homology detection. *Bioinformatics*, *24*(10), 1264–1270.

Damoulas, T., & Girolami, M. A. (2008b). Probabilistic multi-class multikernel learning: On protein fold recognition and remote homology detection. *Bioinformatics*, *24*(10), 1264–1270. https://doi.org/10.1093/bioinformatics/btn112

Desikan, R. S., Segonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into Gyral based regions of interest. *NeuroImage*, *31*(3), 968–980.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 957–980). Boston, MA: Springer.

Duits, F. H., Teunissen, C. E., Bouwman, F. H., Visser, P.-J., Mattsson, N., Zetterberg, H., … van der Flier, W. M. (2014). The cerebrospinal fluid Alzheimer profile: Easily said, but what does it mean? *Alzheimer's & Dementia*, *10*(6), 713–723. https://doi.org/10.1016/j.jalz.2013.12.023

Elahi, S., Bachman, A. H., Lee, S. H., Sidtis, J. J., & Ardekani, B. A. (2015). Corpus callosum atrophy rate in mild cognitive impairment and prodromal Alzheimer's disease. *Journal of Alzheimer's Disease*, *45*(3), 921–931.

Ellis, J. R., Nathan, P. J., Villemagne, V. L., Mulligan, R. S., Saunder, T., Young, K., … Rowe, C. C. (2009). Galantamine-induced improvements in cognitive function are not related to alterations in alpha (4)beta (2) nicotinic receptors in early Alzheimer's disease as measured in vivo by 2[18F]fluoro-A-85380 PET. *Psychopharmacology*, *202*(1–3), 79–91. https://doi.org/10.1007/s00213-008-1347-9

Eskildsen, S. F., Coupe, P., Garcia-Lorenzo, D., Fonov, V., Pruessner, J. C., & Collins, D. L. (2013). Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, *65*, 511–521. https://doi.org/10.1016/j.neuroimage.2012.09.058

Fagan, A. M., Roe, C. M., Xiong, C., Mintun, M. A., Morris, J. C., & Holtzman, D. M. (2007). Cerebrospinal fluid tau/β-amyloid42 ratio as a prediction of cognitive decline in nondemented older adults. *Archives of Neurology*, *64*(3), 343–349. https://doi.org/10.1001/archneur.64.3.noc60123

Fiot, J. B., Raguet, H., Risser, L., Cohen, L. D., Fripp, J., & Vialard, F. X. (2014). Longitudinal deformation models, spatial regularizations and learning strategies to quantify Alzheimer's disease progression. *NeuroImage: Clinical*, *4*, 718–729. https://doi.org/10.1016/j.nicl.2014.02.002

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Fitzpatrick, M., & Sonka, M. (2000). *Handbook of medical imaging vol 2: Medical image processing & analysis (PM80)*. Bellingham, WA: SPIE-International Society for Optical Engineering.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., … Shinohara, R. T. (2018). Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024

Fortin, J.-P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., … Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, *161*, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., & Shinohara, R. T. (2016). Alzheimer's Disease Neuroimaging Initiative, removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, *132*, 198–212. https://doi.org/10.1016/j.neuroimage.2016.02.036

Frisoni, G. B., Boccardi, M., Barkhof, F., Blennow, K., Cappa, S., Chiotis, K., … Winblad, B. (2017). Strategic roadmap for an early diagnosis of Alzheimer's disease based on biomarkers. *The Lancet Neurology*, *16*(8), 661–676. https://doi.org/10.1016/S14744422(17)30159-X

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. A. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 67–77. https://doi.org/10.1038/nrneurol.2009.215

Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology*, *80*(19), 1778–1783.

Huang, M., Yang, W., Feng, Q., Chen, W., & Alzheimer's Disease Neuroimaging Initiative. (2017). Longitudinal measurement and hierarchical classification framework for the prediction of Alzheimer's disease. *Scientific Reports*, *7*, 39880.

Huffman, C., Sobral, H., & Teran-Hinojosa, E. (2019). Laser-induced breakdown spectroscopy spectral feature selection to enhance classification capabilities: A t-test filter approach. *Spectrochimica Acta Part B: Atomic Spectroscopy*, *162*, 105721. https://doi.org/10.1016/j.sab.2019.105721

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., … Silverberg, N. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, *14*(4), 535–562. https://doi.org/10.1016/j.jalz.2018.02.018

Jack, C. R., Petersen, R. C., Xu, Y. C., Waring, S. C., O'Brien, P. C., Tangalos, E. G., … Kokmen, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, *49*(3), 786–794. https://doi.org/10.1212/WNL.49.3.786

Jack, C. R., Wiste, H. J., Weigand, S. D., Therneau, T. M., Lowe, V. J., Knopman, D. S., … Petersen, R. C. (2017). Defining imaging biomarker cut points for brain aging and Alzheimer's disease. *Alzheimer's & Dementia*, *13*(3), 205–216. https://doi.org/10.1016/j.jalz.2016.08.005

Jack, C. R., Jr., Barrio, J. R., & Kepe, V. (2013). Cerebral amyloid PET imaging in Alzheimer's disease. *Acta Neuropathologica*, *126*(5), 643–657. https://doi.org/10.1007/s00401-013-1185-7

Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., … Glahn, D. C. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMADTI Working Group. *NeuroImage*, *81*, 455–469. https://doi.org/10.1016/j.neuroimage.2013.04.061

Jefferson, A. L., Gifford, K. A., Damon, S., Chapman, G. W., Liu, D., Sparling, J., … Salat, D. (2015). Gray & white matter tissue contrast differentiates mild cognitive impairment converters from non-converters. *Brain Imaging and Behavior*, *9*(2), 141–148. https://doi.org/10.1007/s11682-014-9291-2

Khan, W., Westman, E., Jones, N., Wahlund, L.-O., Mecocci, P., Vellas, B., … Simmons, A. (2015). Automated hippocampal subfield measures as predictors of conversion from mild cognitive impairment to Alzheimer's disease in two independent cohorts. *Brain Topography*, *28*(5), 746–759. https://doi.org/10.1007/s10548-014-0415-1

Knopman, D. S., Boeve, B. F., & Petersen, R. C. (2003). Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia. *Mayo Clinic Proceedings*, *78*, 1290–1308.

Kochunov, P., Jahanshad, N., Marcus, D., Winkler, A., Sprooten, E., Nichols, T. E., … van Essen, D. C. (2015). Heritability of fractional anisotropy in human white matter: A comparison of human connectome project and ENIGMA-DTI data. *NeuroImage*, *111*, 300–311. https://doi.org/10.1016/j.neuroimage.2015.02.050

la Joie, R., Perrotin, A., Barre, L., Hommet, C., Mezenge, F., Ibazizene, M., … Chetelat, G. (2012). Region-specific hierarchy between atrophy, hypometabolism, and -amyloid (A) load in Alzheimer's disease dementia. *Journal of Neuroscience*, *32*(46), 16265–16273. https://doi.org/10.1523/JNEUROSCI.2170-12.2012

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., & Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: Biomarker analysis and shared morphometry database. *Scientific Reports*, *8*(1), 1–16. https://doi.org/10.1038/s41598-018-29295-9

Li, Y. D., Dong, H. B., Xie, G. M., & Zhang, L. J. (2013). Discriminative analysis of mild Alzheimer's disease and normal aging using volume of hippocampal subfields and hippocampal mean diffusivity: An in vivo magnetic resonance imaging study. *American Journal of Alzheimer's Disease and Other Dementias*, *28*(6), 627–633. https://doi.org/10.1177/1533317513494452

Liu, F., Zhou, L., Shen, C., & Yin, J. (2013). Multiple kernel learning in the primal for multi-modal Alzheimer's disease classification. *IEEE Journal of Biomedical and Health Informatics*, *18*(3), 984–990. https://doi.org/10.1109/JBHI.2013.2285378

Liu, J., Li, Z., Chen, K., Yao, L., Wang, Z., Li, K., & Guo, X.. *Comparison of gray matter volume and thickness for analysis of cortical changes in Alzheimer's disease*. Proceedings of the SPIE 7965, Medical Imaging 2011: Biomedical Applications in Molecular, Structural, and Functional Imaging, 79652E (15 March 2011). doi:https://doi.org/10.1117/12.877624.

Loew, M. H. (2000). Feature extraction. *Handbook of Medical Imaging*, *2*, 273–342.

Lorenzi, M., Pennec, X., Frisoni, G. B., & Ayache, N. (2014). Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiology of Aging*, *36*, S42–S52. https://doi.org/10.1016/j.neurobiolaging.2014.07.046

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., & Beg, M. F. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Scientific Reports*, *8*(1), 5697. https://doi.org/10.1038/s41598-018-22871-z

Ma, D., Holmes, H. E., Cardoso, M. J., Modat, M., Harrison, I. F., Powell, N. M., … Ourselin, S. (2019). Study the longitudinal in vivo and cross-sectional ex vivo brain volume difference for disease progression and treatment effect on mouse model of tauopathy using automated MRI structural parcellation. *Frontiers in Neuroscience*, *13*, 11.

Ma, D., Popuri, K., Bhalla, M., Sangha, O., Lu, D., Cao, J., … Beg, M. F. (2019). Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database. *Human Brain Mapping*, *40*(5), 1507–1527. https://doi.org/10.1002/hbm.24463

Malone, I. B., Cash, D., Ridgway, G. R., MacManus, D. G., Ourselin, S., Fox, N. C., & Schott, J. M. (2013). MIRIADPublic release of a multiple time point Alzheimer's MR imaging dataset. *NeuroImage*, *70*, 33–36. https://doi.org/10.1016/j.neuroimage.2012.12.044

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, *19*, 1498–1507.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, *19*(9), 1498–1507.

Maruyama, M., Arai, H., Sugita, M., Tanji, H., Higuchi, M., Okamura, N., … Sasaki, H. (2001). Cerebrospinal fluid amyloid β142 levels in the mild cognitive impairment stage of Alzheimer's disease. *Experimental Neurology*, *172*(2), 433–436. https://doi.org/10.1006/exnr.2001.7814

Min, J., Moon, W. J., Jeon, J. Y., Choi, J. W., Moon, Y. S., & Han, S. H. (2017). Diagnostic efficacy of structural MRI in patients with mild-to-moderate Alzheimer disease: Automated volumetric assessment versus visual assessment. *American Journal of Roentgenology*, *208*(3), 617–623. https://doi.org/10.2214/AJR.16.16894

Mitchell, A. J., & Shiri-Feshki, M. (2008). Temporal trends in the long term risk of progression of mild cognitive impairment: A pooled analysis. *Journal of Neurology, Neurosurgery and Psychiatry*, *79*(12), 1386–1391. https://doi.org/10.1136/jnnp.2007.142679

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01* (pp. 841–848). Cambridge, MA: MIT Press Retrieved from http://dl.acm.org/citation.cfm?id=2980539.2980648

O'Brien, P. C., & Dyck, P. J. (1995). Procedures for setting normal values. *Neurology*, *45*, 17–23. https://doi.org/10.1212/WNL.45.1.17

Ossenkoppele, R., Jansen, W. J., Rabinovici, G. D., Knol, D. L., van der Flier, W. M., van Berckel, B. N., … Brooks, D. J. (2015). Prevalence of amyloid PET positivity in dementia syndromes: A meta-analysis. *JAMA*, *313*(19), 1939–1949. https://doi.org/10.1001/jama.2015.4669

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., … Weiner, M. W. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/WNL.0b013e3181cb3e25

Popuri, K., Balachandar, R., Alpert, K., Lu, D., Bhalla, M., Mackenzie, I. R., … Beg, M. F. (2018). Development and validation of a novel dementia of Alzheimer's type (DAT) score based on metabolism FDG-PET imaging. *NeuroImage: Clinical*, *18*, 802–813. https://doi.org/10.1016/j.nicl.2018.03.007

Potvin, O., Dieumegarde, L., & Duchesne, S. (2017). Freesurfer cortical normative data for adults using Desikan-Killiany-Tourville and ex vivo protocols. *NeuroImage*, *156*, 43–64. https://doi.org/10.1016/j.neuroimage.2017.04.035

Potvin, O., Mouiha, A., Dieumegarde, L., & Duchesne, S. (2016). Normative data for subcortical regional volumes over the lifetime of the adult human brain. *NeuroImage*, *137*, 9–20. https://doi.org/10.1016/j.neuroimage.2016.05.016

Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F., & Dickerson, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research: Neuroimaging*, *194*(1), 7–13. https://doi.org/10.1016/j.pscychresns.2011.06.014

Prince, M., Comas-Herrera, A., Knapp, M., Guerchet, M., & Karagiannidou, M. (2016). *World Alzheimer report 2016 improving healthcare for people living with dementia coverage, quality and costs now and in the future, Technical Report*. London: Alzheimer's Disease International.

Raamana, P. R., Weiner, M. W., Wang, L., & Beg, M. F. (2015). Alzheimer's Disease Neuroimaging Initiative, thickness network features for prognostic applications in dementia. *Neurobiology of Aging*, *36*, S91–S102. https://doi.org/10.1016/j.neurobiolaging.2014.05.040

Raji, C. A., Lopez, O. L., Kuller, L. H., Carmichael, O. T., & Becker, J. T. (2009). Age, Alzheimer disease, and brain structure. *Neurology*, *73*(22), 1899–1905. https://doi.org/10.1212/WNL.0b013e3181c3f293

Ritchie, C., Smailagic, N., Noel-Storr, A. H., Ukoumunne, O., Ladds, E. C., & Martin, S. (2017). CSF tau and the CSF tau/ABeta ratio for the diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). *Cochrane Database of Systematic Reviews*, *3*, CD010803.

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., … Davatzikos, C. (2017). Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophrenia Bulletin*, *44*(5), 1035–1044. https://doi.org/10.1093/schbul/sbx137

Rusinek, H., de Leon, M. J., George, A. E., Stylopoulos, L. A., Chandra, R., Smith, G., … Kowalski, H. (1991). Alzheimer disease: Measuring loss of cerebral gray matter with MR imaging. *Radiology*, *178*(1), 109–114. https://doi.org/10.1148/radiology.178.1.1984287

Schmitter, D., Roche, A., Marechal, B., Ribes, D., Abdulkadir, A., Bach-Cuadra, M., … Krueger, G. (2015). An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease. *NeuroImage: Clinical*, *7*(1), 7–17. https://doi.org/10.1016/j.nicl.2014.11.001

Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels*. Cambridge, MA: The MIT Press.

Schwarz, C. G., Gunter, J. L., Wiste, H. J., Przybelski, S. A., Weigand, S. D., Ward, C. P., … Jack, C. R. (2016). A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity. *NeuroImage: Clinical*, *11*, 802–812. https://doi.org/10.1016/j.nicl.2016.05.017

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Shi, F., Liu, B., Zhou, Y., Yu, C., & Jiang, T. (2009). Hippocampal volume and asymmetry in mild cognitive impairment and Alzheimer's disease: Metaanalyses of MRI studies. *Hippocampus*, *19*, 1055–1064. https://doi.org/10.1002/hipo.20573

Silbert, L. C., Quinn, J. F., Moore, M. M., Corbridge, E., Ball, M. J., Murdoch, G., … Kaye, J. A. (2003). Changes in premorbid brain volume predict Alzheimer's disease pathology. *Neurology*, *61*(4), 487–492. https://doi.org/10.1212/01.WNL.0000079053.77227.14

Sørensen, L., Igel, C., Liv Hansen, N., Osler, M., Lauritzen, M., Rostrup, E., & Nielsen, M. (2016). Alzheimer's Disease Neuroimaging Initiative and the Australian imaging biomarkers and lifestyle flagship study of ageing, early detection of Alzheimer's disease using MRI hippocampal texture. *Human Brain Mapping*, *37*(3), 1148–1161. https://doi.org/10.1002/hbm.23091

Stout, J. C., Jernigan, T. L., Archibald, S. L., & Salmon, D. P. (1996). Association of dementia severity with cortical gray matter and abnormal white matter volumes in dementia of the Alzheimer type. *Archives of Neurology*, *53*(8), 742–749. https://doi.org/10.1001/archneur.1996.00550080056013

Suk, H. I., Lee, S. W., & Shen, D. (2017). Deep ensemble learning of sparse regression models for brain disease diagnosis. *Medical Image Analysis*, *37*, 101–113. https://doi.org/10.1016/j.media.2017.01.008

Sun, Z., van de Giessen, M., Lelieveldt, B. P. F., & Staring, M. (2017). Detection of conversion from mild cognitive impairment to Alzheimer's disease using longitudinal brain MRI. *Frontiers in Neuroinformatics*, *11*, 16. https://doi.org/10.3389/fninf.2017.00016

Tapiola, T., Alafuzoff, I., Herukka, S.-K., Parkkinen, L., Hartikainen, P., Soininen, H., & Pirttila, T. A. (2009). Cerebrospinal fluid β-amyloid 42 and tau proteins as biomarkers of Alzheimer-type pathologic changes in the brain. *Archives of Neurology*, *66*(3), 382–389. https://doi.org/10.1001/archneurol.2008.596

Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., … Ye, J. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*, *145*, 389–408. https://doi.org/10.1016/j.neuroimage.2015.11.057

Thompson, P. M., Hayashi, K. M., de Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., … Toga, A. W. (2003). Dynamics of gray matter loss in Alzheimer's disease. *The Journal of Neuroscience*, *23*(3), 994–1005.

Thompson, P. M., Hayashi, K. M., Sowell, E. R., Gogtay, N., Giedd, J. N., Rapoport, J. L., … Toga, A. W. (2004). Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia. *Neuroimage*, *23*(Suppl 1), S2–S18.

Toledo, J. B., Zetterberg, H., Van Harten, A. C., Glodzik, L., MartinezLage, P., Bocchio-Chiavetto, L., … Trojanowski, J. Q. (2015). Alzheimer's disease cerebrospinal fluid biomarker in cognitively normal subjects. *Brain*, *138*(9), 2701–2715. https://doi.org/10.1093/brain/awv199

Varol, E., Gaonkar, B., Erus, G., Schultz, R., & Davatzikos, C.. *Feature ranking based nested support vector machine ensemble for medical image classification*. Proceedings of the IEEE International Symposium on Biomedical Imaging, 2012. pp. 146–149. doi: https://doi.org/10.1109/ISBI.2012.6235505.

Visser, P. J., Verhey, F. R., Boada, M., Bullock, R., De Deyn, P. P., Frisoni, G. B., … Winblad, B. (2008). Development of screening guidelines and clinical criteria for predementia Alzheimer's disease. *Neuroepidemiology*, *30*(4), 254–265. https://doi.org/10.1159/000135644

Wachinger, C., Salat, D. H., Weiner, M., & Reuter, M. (2016). Whole-brain analysis reveals increased neuroanatomical asymmetries in dementia for hippocampus and amygdala. *Brain*, *139*(Pt 12), 3253–3266. https://doi.org/10.1093/brain/aww243

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). T-test feature selection approach based on term frequency for text categorization. *Pattern Recognition Letters*, *45*, 1–10. https://doi.org/10.1016/j.patrec.2014.02.013

Wang, L., Heywood, A., Stocks, J., Bae, J., Ma, D., Popuri, K., … Rosen, H. (2019). Grant report on PREDICT-ADFTD: Multimodal imaging prediction of AD/FTD and differential diagnosis. *Journal of Psychiatry and Brain Science*, *4*, e190017.

Wang, R. a. (2012). AdaBoost for feature selection, classification and its relation with SVM, a review. *Physics Procedia*, *25*, 800–807. https://doi.org/10.1016/j.phpro.2012.03.160

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., … Trojanowski, J. Q. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, *13*(4), e1–e85. https://doi.org/10.1016/j.jalz.2016.11.007

Yang, C., Zhong, S., Zhou, X., Wei, L., Wang, L., & Nie, S. (2017). The abnormality of topological asymmetry between hemispheric brain white matter networks in Alzheimer's disease and mild cognitive impairment. *Frontiers in Aging Neuroscience*, *9*, 261. https://doi.org/10.3389/fnagi.2017.00261

Yi, H.-A., Moller, C., Dieleman, N., Bouwman, F. H., Barkhof, F., Scheltens, P., … Vrenken, H. (2016). Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(4), 425–432. https://doi.org/10.1136/jnnp-2014-309105

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., & Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, *2*, 735–745. https://doi.org/10.1016/j.nicl.2013.05.004

Yu, M., Linn, K. A., Cook, P. A., Phillips, M. L., McInnis, M., Fava, M., … Sheline, Y. I. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human Brain Mapping*, *39*(11), 4213–4227. https://doi.org/10.1002/hbm.24241

Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895–907. https://doi.org/10.1016/j.neuroimage.2011.09.069

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zhou, N., & Wang, L. (2007). A modified T-test feature selection method and its application on the HapMap genotype data. *Genomics, Proteomics & Bioinformatics*, *5*(3–4), 242–249. https://doi.org/10.1016/S16720229(08)60011-X

Zhu, X., Suk, H. I., Wang, L., Lee, S. W., & Shen, D. (2017). A novel relational regularization feature selection method for joint regression and classification in AD diagnosis. *Medical Image Analysis*, *38*, 205–214. https://doi.org/10.1016/j.media.2015.10.008

How to cite this article: Popuri K, Ma D, Wang L, Beg MF. 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*. 2020;41:4127–4147. https://doi.org/10.1002/hbm.25115