Neurobiology of Aging 59 (2017) 80–90

Contents lists available at ScienceDirect

Neurobiology of Aging

journal homepage: www.elsevier.com/locate/neuaging

Identifying incipient dementia individuals using machine learning and amyloid imaging

Sulantha Mathotaarachchi [a,b], Tharick A. Pascoal [a], Monica Shin [a], Andrea L. Benedet [a], Min Su Kang [a], Thomas Beaudry [a], Vladimir S. Fonov [b], Serge Gauthier [c,d,e,f], Pedro Rosa-Neto [a,b,c,d,e,f,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Translational Neuroimaging Laboratory, McGill University Research Centre for Studies in Aging (MCSA), Douglas Research Institute, McGill University, Montreal, Quebec, Canada
[b] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada
[c] McGill University Research Centre for Studies in Aging (MCSA), Douglas Research Institute, McGill University, Montreal, Quebec, Canada
[d] Douglas Research Institute, McGill University, Montreal, Quebec, Canada
[e] Department of Psychiatry, McGill University, Montreal, Quebec, Canada
[f] Department of Neurology & Neurosurgery, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Received 6 March 2017
Received in revised form 20 June 2017
Accepted 30 June 2017
Available online 11 July 2017

Keywords:
Alzheimer's disease
Mild cognitive impairment
Prediction
Amyloid
Random forest
Random under sampling

ABSTRACT

Identifying individuals destined to develop Alzheimer's dementia within time frames acceptable for clinical trials constitutes an important challenge to design studies to test emerging disease-modifying therapies. Although amyloid-β protein is the core pathologic feature of Alzheimer's disease, biomarkers of neuronal degeneration are the only ones believed to provide satisfactory predictions of clinical progression within short time frames. Here, we propose a machine learning–based probabilistic method designed to assess the progression to dementia within 24 months, based on the regional information from a single amyloid positron emission tomography scan. Importantly, the proposed method was designed to overcome the inherent adverse imbalance proportions between stable and progressive mild cognitive impairment individuals within a short observation period. The novel algorithm obtained an accuracy of 84% and an under-receiver operating characteristic curve of 0.91, outperforming the existing algorithms using the same biomarker measures and previous studies using multiple biomarker modalities. With its high accuracy, this algorithm has immediate applications for population enrichment in clinical trials designed to test disease-modifying therapies aiming to mitigate the progression to Alzheimer's disease dementia.

1. Introduction

Alzheimer's disease is the most common form of dementia, accounting for approximately 60%–80% of the cases with alarming economic and social costs (Pouryamout et al., 2012). It is well understood that Alzheimer's disease pathophysiology follows a gradual course with pathologies developing over decades before symptom onset (Braak and Braak, 1991; Mosconi et al., 2007). Thus, there is an increased interest to advance the disease-modifying clinical trials to the predementia stages of Alzheimer's disease given the better chances of obtaining tangible disease-modifying effects. This has been encouraged by the identification of biomarkers to detect the pathologic signatures at the early stages of the disease (Buerger et al., 2006; Fagan et al., 2006; McEvoy and Brewer, 2010; Morris et al., 2009). However, given the high prevalence of cognitively normal elderly individuals carrying the Alzheimer's disease pathophysiology, accurately identifying individuals who are in the early stages of Alzheimer's disease has been a significant challenge. Thus, the development of effective clinical trials has been severely hampered by not being able to recruit individuals who have a higher likelihood of developing Alzheimer's disease (Holland et al., 2012) within a time period acceptable to conduct a study.

* Corresponding author at: Translational Neuroimaging Laboratory, McGill University Research Centre for Studies in Aging, 6875 La Salle Blvd - FBC room 3149, Montreal, Québec H4H 1R3, Canada. Tel.: +1 514 766-2010; fax: +1 514 888-4050.
E-mail address: pedro.rosa@mcgill.ca (P. Rosa-Neto).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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.

http://dx.doi.org/10.1016/j.neurobiolaging.2017.06.027

Identification of effective biomarker signatures to determine the progression of dementia has been a significant challenge to accurately predict the short-term clinical progression. It is believed that the biomarkers of neurodegeneration such as glucose metabolism and brain atrophy would be the most effective, as the other pathologic markers such as brain amyloid-β (Aβ) deposition and formation of hyperphosphorylated tau tangles occur years or decades before the symptom onset (Jack et al., 2010a,b, 2013). However, numerous studies based on neurodegenerative biomarkers have indicated a decrease in the ability to predict the clinical progression when the observation time is more than 18 months. Given that the accumulation of Aβ in the brain is a hallmark characteristic of Alzheimer's disease, amyloid-based biomarkers can be assumed to provide a better sensitivity to Alzheimer's disease pathology over a longer observation period (Trzepacz et al., 2014).

Studies conducted to evaluate the utility of amyloid biomarkers have used the cerebrospinal fluid (CSF) Aβ measurements due to the relatively higher availability compared with amyloid positron emission tomography (PET) biomarkers. These studies have often combined other CSF measurements such as tau and p-tau concentrations and magnetic resonance imaging (MRI) biomarkers (Apostolova et al., 2014; Davatzikos et al., 2011; Hall et al., 2015; Yang et al., 2012; Young et al., 2012) to accommodate for the limited dimensionality offered by the CSF measurements. However, the possibility to regionally assess Alzheimer's disease pathophysiology and the increased availability of amyloid PET images in large-scale aging studies have enabled the evaluation of this imaging biomarker as a diagnostic tool.

As the earliest detectable clinical stage of the trajectory toward dementia, mild cognitive impairment (MCI) provides an attractive point of entry to disease-modifying interventions (Markesbery, 2010). However, not all MCI individuals carry Alzheimer's disease pathophysiology and progress to Alzheimer's disease in an optimal time frame for clinical trial. Thus, identification of biomarker signatures of MCI individuals on the verge of progression to dementia and predicting the likelihood of conversion has immediate application in disease-modifying clinical trials by including those individuals with high likelihood to progress. This will increase the statistical power of the clinical trial by reducing false positives and would directly impact the associated costs. Furthermore, these information would also provide important prognostic information for the patients and their families to plan and manage treatment and care (Gelosa and Brooks, 2012).

The prediction of the clinical progression to dementia from MCI phase constitutes a challenge due to the considerably lower number of progressive MCI (pMCI) individuals compared with the stable MCI (sMCI) individuals, invariably creating imbalanced classes (pMCI and sMCI). In a classification study, if the number of samples from a class (sMCI) greatly outnumbers the other classes (pMCI), the classifier tends to favor the over-represented majority class (sMCI) by ignoring the incorrect prediction in the minority class (pMCI). This constitutes a challenge known as the "class imbalance problem" and leads to decreased sensitivity of the prediction, which has been currently overlooked in Alzheimer's disease research. While there are a number of techniques developed to address this challenge, such as class weight adjusting, oversampling the minority class, and undersampling the majority class (Japkowicz and Stephen, 2002), the latter is identified to perform the best (Seiffert et al., 2008, 2010). Random under sampling (RUS) is a method of under sampling, which will train each of its base learners with the full minority class and a randomly under sampled majority class. By doing this, each base learner is trained with a balanced set of samples from both pMCI and sMCI classes, removing the bias toward any particular class.

Amyloid PET images used in the study are [^{18}F]Florbetapir PET scans acquired from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The study cohort presented a pMCI rate of 15.75%, similar to the rates presented by other cohort studies and community studies (Mitchell and Shiri-Feshki, 2009). This presents the aforementioned class-imbalanced challenge. Current study utilizes RUS to overcome this challenge. The novel prediction algorithm is presented as RUS random forest (RUSRF) and is used to evaluate the utility of amyloid PET as a diagnostic tool.

The present study aims to evaluate the utility of amyloid PET imaging as an early detection tool for individuals in the Alzheimer's disease pathway. Specifically, we aim to predict if an MCI individual will progress to Alzheimer's disease or remain MCI within a time period of 24 months based on the amyloid PET measurements at baseline. We further believe that the proposed technique can be used to complement the multibiomarker enrichment strategies for clinical trials (Wolz et al., 2016).

This novel technique capitalizes on an optimized feature extraction method for amyloid PET, and a method for addressing the class imbalance problem in the context of a random forest machine learning classifier (Breiman, 2001). We further compare the performance of RUSRF with other widely-used prediction algorithms such as the Support Vector Machine (SVM), L1 and L2 regularized logistic regression, and random forest. We believe that the novel RUSRF algorithm will outperform these widely-used prediction algorithms and that the amyloid PET biomarker might serve as an important early diagnostic tool in both research and clinical environments.

2. Materials and methods

2.1. Subjects and data acquisition

Data used in the preparation of this article were obtained from the ADNI database. ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early Alzheimer's disease.

We selected 273 ADNI-GO/2 participants meeting the criteria for single-domain or multidomain amnestic MCI and acquired their [^{18}F] Florbetapir PET, demographic, APOE4 genotype information and the clinical diagnosis with the corresponding diagnosis information for the subsequent 24-month follow-up visit (mean follow-up time: 24.37 months, standard deviation: 1.51 months). For PET image processing purposes, the matching T1-weighted MRI images were also obtained. The eligibility criteria of MCI were participants with mini-mental state examination (MMSE) scores equal to or greater than 24, a clinical dementia rating of 0.5, subjective and objective memory loss, normal activities of daily living, and the absence of other neuropsychiatric disorders including Alzheimer's disease (Petersen, 2003).

The subjects were categorized into 2 groups, sMCI and pMCI, based on the diagnosis on their follow-up visit. If a subject was diagnosed to be Alzheimer's disease in the follow-up visit, that subject was categorized as pMCI, and if the subject remained MCI at the 24-month follow-up visit, he or she was categorized as sMCI. Of the 273 individuals used in the study, only 43 (15.75%) individuals were diagnosed as probable Alzheimer's disease in their follow-up visit and were categorized as pMCI. Demographical information of the sample of individuals used in the study is presented in Table 1.

2.2. Image processing

The acquired T1-weighted MRI images were processed using the CIVET image processing pipeline (Ad-Dab'bagh et al., 2006; Zijdenbos et al., 2002), where the images underwent non-

Table 1
Demographics of subjects included in the study and their classification in training and testing sets

Characteristic	sMCI	pMCI
N (273)	230	43
Age (SD)	71.45 (7.52)	73.25 (7.65)
Males (females)	135 (95)	23 (20)
Education(SD)	16.2 (2.71)	15.86 (2.81)
APOE 4−positive (Neg)***	81 (149)	33 (10)
MMSE (SD)***	28.22 (1.64)	26.91 (1.87)
Training (S_{Train}) (pMCI ratio)	161	30 (15.71%)
Testing (S_{Test}) (pMCI ratio)	69	13 (15.85%)

Key: MMSE, mini-mental state examination; pMCI, progressive MCI.
***$p < 0.05$.

uniformity correction (Sled et al., 1998), brain masking (Smith, 2002), linear and nonlinear registrations to the ICBM 152 Template (Fonov et al., 2009; Mazziotta et al., 1995) using the ANIMAL (Collins and Evans, 1997; Collins et al., 1995; Robbins et al., 2004) image registration algorithm. Downloaded [^{18}F]Florbetapir PET images were already preprocessed to perform motion correction and to acquire uniform spatial resolution according to the steps outlined in (online: *http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/*, Accessed: 28-03-2016). The PET images were then registered to the subjects' own T1-weighted MRI images and were spatially normalized to the ICBM 152 template. They were then normalized for the regional intensities from the cerebellum gray matter and the global white matter to generate [^{18}F]Florbetapir PET standard uptake value ratio (SUVR) images.

2.3. Subject classification

Two sets of subjects were generated from the total set of individuals as training set (S_{Train}) and testing set (S_{Test}; Table 1). The training set was used in the feature selection step and training the machine learning algorithm, while the testing set was used solely to evaluate the performance of the trained machine learning algorithm. Since the testing set is not used in the feature selection step or in training, we void any bias or adverse effect from "double dipping" in the prediction of progression to Alzheimer's disease. The ratio between the sMCI and pMCI individuals was preserved in both the training set and the testing set.

2.4. Feature selection

To identify the brain regions from where the PET SUVR values are read as features for the predictors, we employed a voxel-wise logistic regression analysis using the VoxelStats (Mathotaarachchi et al., 2016) toolbox. For each brain voxel, the following logistic function was solved:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * [^{18}F]Florbetapir + \beta_2 * Age + \beta_3 * Gender + \beta_4 * APOE4$$

where π = probability of progressing to Alzheimer's disease in 24 months. Then, for each voxel, the scaled increase in odds (scaled odds ratio) for one standard deviation (σ) increase in [^{18}F]Florbetapir PET SUVR was calculated as follows:

$$\text{Odds Ratio}_{Scaled}(\text{OR}_{Scaled}) = e^{\beta_1 * \sigma}$$

Anatomically significant brain regions with $\text{OR}_{\text{scaled}} > 1.5$ were selected based on expert opinion, and PET SUVR intensities from these regions were extracted as the mean of a $3 \times 3 \times 3$ cluster of voxels. The above regional PET SUVR intensities along with age, gender, and APOE4 genotype status were fed into the predictors as features.

2.5. Prediction

Prediction was performed using the novel algorithm RUSRF based on the random forest classifier (Breiman, 2001). The novel algorithm uses sampling techniques to train multiple random forest predictors to calculate the final prediction. The primary goal of the present application is to predict progression to Alzheimer's disease; however, the population contains only 15.71% pMCI individuals. This constitutes a class imbalance problem with a majority sMCI class. In this case, the predictors tend to be biased to predict the majority class (sMCI) and to ignore the misclassification in the minority class (pMCI) (Yen and Lee, 2009). To overcome this, the novel algorithm utilizes a data sampling technique named "random undersampling," where the majority class is randomly sampled to create a pseudo training set with equal proportions in all the classes. This technique is closely related to the RUSBoost technique (Seiffert et al., 2008); however, the base classifier used in the present study is random forest and the iterative boosting method is not employed as the random forest classifier uses feature bagging to achieve independence and randomness.

In summary, RUSRF algorithm will train K number of random forest predictors with individual pseudo training sets with equal class distributions. The number of minority class samples (N^{α}) in a pseudo training set is equal to the total number of minority class samples in the original training set. Thus, all K random forest predictors are trained for all the minority class instances. The number of majority class samples is calculated to be a fraction (f) higher than the number of samples in the minority class. The final probability prediction is calculated as the mean probability across all the K random forests, and the final prediction is calculated as the final probability prediction greater than 0.5.

The implementation of the RUSRF algorithm was done in Python programming language using the random forest implementation in the scikit-learn library (Online: *http://scikit-learn.org/stable/*, Accessed: 28-04-2016).

2.6. Performance evaluation and comparison

The RUSRF algorithm has 2 primary parameters to be optimized based on the application, which are the number of random forest predictors and the number of decision trees in each of the random forest predictor. To estimate the best combination of these 2 parameters, we employed a nested cross-validation technique (Huttunen et al., 2012; Moradi et al., 2015). The outer cross-validation step divided the training set (S_{Train}) in to 10 subsets ($S^i_{Train}; i = 1 \ldots 10$) preserving the class ratios of the original population (stratified cross validation). In each training step i, S^i_{Train} was subdivided (using stratified cross validation) into 10 subsets ($S^{ij}_{Train}; j = 1 \ldots 10$) and was passed to RUSRF algorithm with a tuple of parameters (number of random forest predictors, number of trees in each predictor) selected from a grid of parameter combinations for training. The best set of parameters was selected based on the AUC from the internal 10-fold cross validation. The performance of the predictor was then evaluated based on the AUC, accuracy (number of correct classifications divided by the total number of samples [in S_{Test}]), F1 score (harmonic mean of precision and recall), sensitivity (number of correctly classified pMCI subjects divided by the total pMCI subjects), and specificity (number of correctly classified sMCI subjects divided by the total sMCI subjects) for the original test set (S_{Test}). These values were calculated based on the predictions of each outer cross-validation step i, on the testing set (S_{Test}). By evaluating the performance based on the original test set (S_{Test}), we recorded the true generalization ability of the RUSRF predictor.

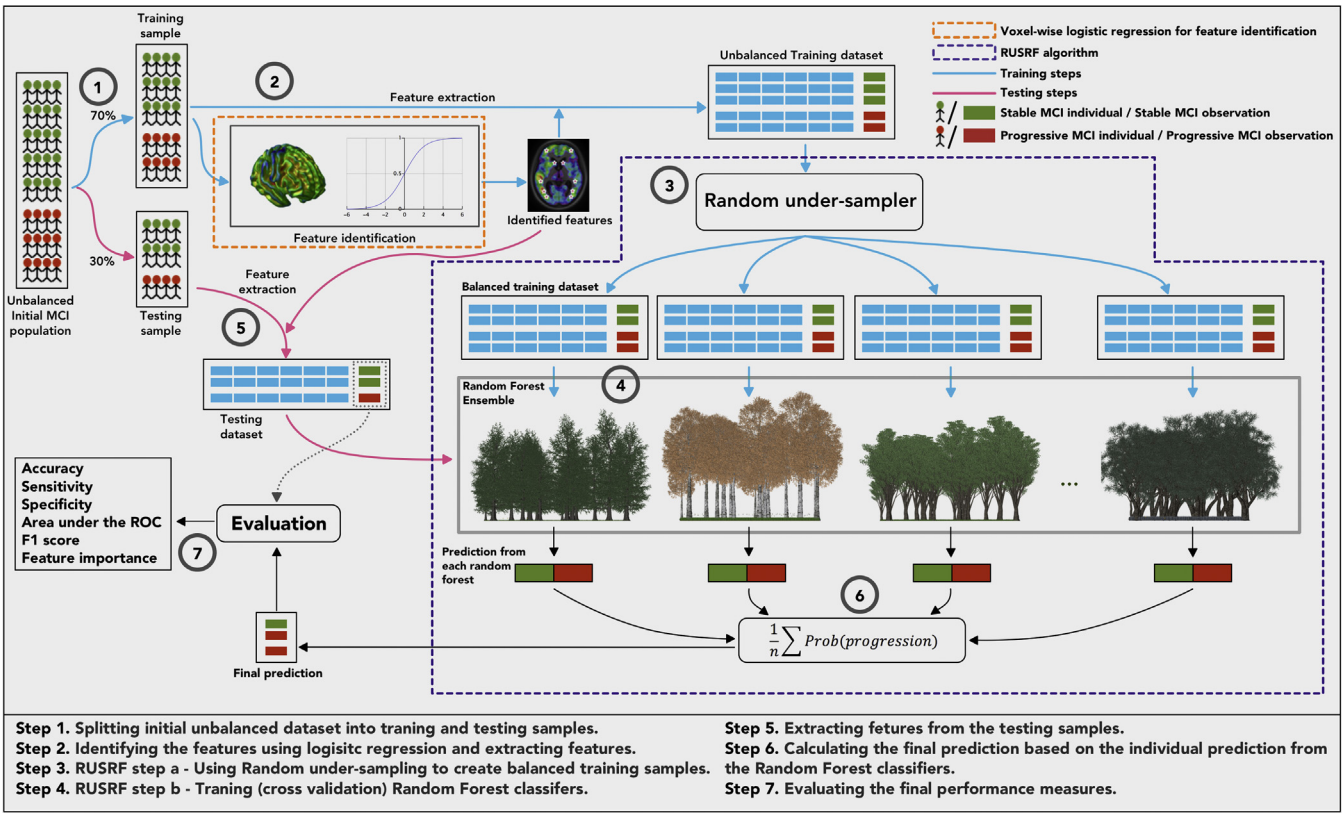

Fig. 1. Summary of the steps followed to train the RUSRF algorithm to predict the progression of dementia. Abbreviations: MCI, mild cognitive impairment; ROC, receiver operating characteristic; RUSRF, RUS random forest.

To examine the performance gain from the RUSRF predictor, we trained 4 other widely-used classifiers such as SVM, logistic regression with L1 regularization, logistic regression with L2 regularization, and random forest (Breiman, 2001) using the same set of features and compared with the performance of RUSRF. Evaluating the optimum parameters and the performance for these widely-used classifiers were also done using a nested cross-validation as described above. McNemar's χ^2 test was performed to compare each of the classifiers against random predictors. The random predictors were developed by training each of the classification algorithms with a randomly permuted training set to disrupt the relationships between the class labels and input features. Furthermore, a paired t-test was performed to compare AUC and F1 scores of the RUSRF predictor against the other 4 predictors, while McNemar's χ^2 test was used to compare the sensitivity and specificity. However, the overall accuracy of each predictor was not used as a measure of comparison because it can misinterpret the performance when the sample has imbalanced class probabilities. The resulting p values were Bonferroni adjusted to correct for multiple comparisons. The classification analysis was carried out in Python programming language using the scikit-learn library and the statistical analysis was carried out using the R programming language.

In a binary classification, the random forest algorithm performs implicit feature selection based on the parameter called "Gini importance" (Breiman, 2001), which is calculated as the accumulated decrease in "Gini impurity $i(t)$". This is defined as; $i(t) = 1 - p_0^2 - p_1^2$. Here $p_k = \frac{n_k}{n}$, which is the fraction of the n_k samples from class $k = \{0, 1\}$ out of the total n samples at a node t in a binary decision tree in random forest. The decrease in Gini impurity results from splitting the node by a threshold τ_α on variable α, and sending the samples to 2 subnodes, t_l and t_r, and is calculated as $\Delta i(t) = i(t) - p_l i(t_l) - p_r i(t_r)$, where $p_l = \frac{n_l}{n}$ and $p_r = \frac{n_r}{n}$, the relative sample fractions at nodes t_l and t_r. The optimum split is then determined, which results in the maximum decrease in Gini impurity at node t. The accumulation of these decreases in Gini impurity for all nodes t in all trees T in the Forest, is known as the Gini importance (I_G) for variable α.

$$I_G(\alpha) = \sum_T \sum_t \Delta i_\alpha(t, T)$$

This measure indicates how often a variable is used for a split and how large its discriminative power is in the classification study. In the RUSRF algorithm, the mean of this measure across all the random forests, K, is used as the measure of feature importance for variable α.

$$\text{Importance}(\alpha) = \frac{1}{K} \sum_K \sum_T \sum_t \Delta i_\alpha(t, T, K)$$

Steps followed in the present study are summarized in Fig. 1.

3. Results

Demographic information of the individuals included in this study is summarized in Table 1. Of the 273 individuals, 43 (15.75%) were diagnosed with probable Alzheimer's disease after the 24-month follow-up and were categorized as pMCI.

Voxel-wise logistic regression analysis resulted in a scaled odds ratio map with the highest values in the orbitofrontal cortex, mid frontal cortex, mid temporal cortex, temporal occipital junction, posterior cingulate cortex, angular gyrus, precuneus, putamen, and

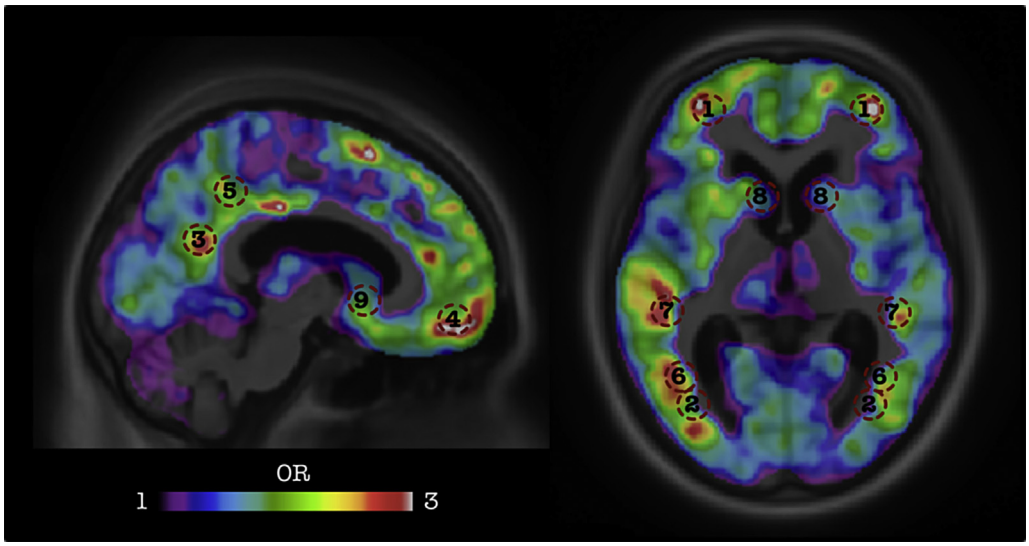

Fig. 2. Regions with the highest odds ratio values. (1) mid frontal cortex, (2) angular gyrus, (3) posterior cingulate cortex, (4) orbitofrontal cortex, (5) precuneus, (6) temporal occipital junction, (7) mid temporal cortex, (8) Putamen, and (9) nucleus accumbens are indicated.

the nucleus accumbens (Fig. 2). The SUVR values from bilateral brain regions were extracted as described above and subsequently averaged to be used as regional SUVRs. These regional SUVR values were then used as features on the predictors together with age, gender, and APOE4 gene status.

Predictive performances from the 5 predictors are shown in Table 2. The RUSRF predictor has the highest predictive performance for both sMCI and pMCI individuals based on the average area under the receiver operating characteristic curve (AUC) (0.906 [±0.009]) and F1 score (0.583 [±0.044]). The RUSRF AUC score is significantly different from both the SVM and L2 logistic regression at $p < 0.0025$ and from random forest at $p < 0.0125$. Furthermore, the RUSRF F1 score is significantly different from SVM at $p < 0.0025$ and from L2 logistic regression at $p < 0.0125$. The sensitivity of the RUSRF predictor only differed from the SVM predictor at $p < 0.025$. The RUSRF specificity was different from L2 logistic regression at $p < 0.0125$ and from SVM, L1 logistic regression, and random forest at $p < 0.025$. The p value thresholds used here for statistical significance were Bonferroni adjusted for multiple comparisons.

According to McNemar's χ^2 test (Table 2), only the RUSRF predictor showed a statistically significant difference with the random predictor at $p < 0.01$. Both regularized logistic regression predictors were significantly different at $p < 0.1$. However, the random forest predictor and the SVM predictor were not significantly different from the random predictor. The χ^2 statistic for the SVM predictor cannot be calculated as the output from the predictions from both SVM and the random predictor were identical. The SVM prediction for each test case, which was calculated as the mean of the prediction from all 10 iterations was 0. This resulted in an exaggerated accuracy of 84.15% (all true 0 cases correctly predicted—specificity of 100%).

Fig. 3A shows the receiver operating characteristic (ROC) curves for the 4 predictors with their respective AUC values. The ROC curve for the novel RUSRF predictor dominated the other curves with an area of 0.91, followed by the L1 logistic predictor, random forest predictor, and the L2 logistic predictor. The AUC <0.5 for the SVM predictor (Table 2) indicated that the SVM predictor is a predictor for the sMCI class but not for the pMCI class.

The relative feature importance from the novel RUSRF predictor is shown in Fig. 3B. The most important features indicated the brain regions with the highest contribution to predict the pMCI class. Aβ accumulations measured by [^{18}F]Florbetapir PET SUVR in the temporal occipital junction, mid temporal cortex, mid frontal cortex, and precuneus were the most important features followed by the accumulation in other brain regions and age, gender, and APOE4 gene status. The feature importance from the random forest classifier indicated the same 4 regions to be the most important features (Supplementary Table 1).

4. Discussion

We have presented a novel algorithm for predicting the progression from MCI to Alzheimer's disease dementia using baseline amyloid PET measurements. The novel RUSRF algorithm uses the random forest classifier as its base classifier along with advanced data sampling techniques to overcome inherent problems of the data sample to improve the prediction accuracy. This algorithm was shown to provide advantages as demonstrated by higher AUC and F1 score when compared with the other widely-used prediction algorithms such as SVM, regularized logistic regression, and random forest. Furthermore, our preliminary data suggested, that this algorithm could perform comparably to the methods employed using MRI, FDG PET, CSF biomarkers, and their combinations (Table 3). It is important to emphasize that the evaluation of the prediction algorithms, employed a completely independent testing sample which was never used in training or feature selection steps. This method of evaluation resembles the expected outcomes in an application (clinical and research) environment, which can be considered encouraging to be used for the early diagnosis of Alzheimer's disease.

Using the RUSRF algorithm, we obtained AUC and F1 scores of 0.906 (0.009) and 0.583 (0.044), respectively, for the independent testing set. Generally, the progression to Alzheimer's disease dementia of an MCI population within 24 months lies around 16.2% (Mitchell and Shiri-Feshki, 2009). Thus, in a prediction study, the pMCI class (the subset of individuals progressing to Alzheimer's disease) is expected to lie close to the same proportion, creating imbalanced classes. If the methods to overcome this challenge is not involved, the prediction algorithm will ignore the errors in the minority class (Japkowicz and Stephen, 2002). This can result in exaggerated accuracies, as the predictor can be biased toward the majority class, but it can also result in lower sensitivity for the

Table 2
Summary of the performance of the predictors used. The statistical significance is measured compared with the RUSRF predictor

Predictor	AUC	Accuracy	F1 score	Sensitivity	Specificity	vs. random predictor McNemar's χ^2 p
SVM	0.436 (0.24)***	0.834 (0.023)	0.091 (0.193)***	0.1 (0.212)*	0.972 (0.061)*	NA
L1 logistic	0.874 (0.047)	0.752 (0.043)	0.523 (0.051)	0.846 (0.051)	0.735 (0.046)*	$8.1990e^{-02*}$
L2 logistic	0.871 (0.007)***	0.751 (0.02)	0.522 (0.023)**	0.854 (0.024)	0.732 (0.023)**	$6.7250e^{-02*}$
Random forest	0.872 (0.03)**	0.89 (0.014)	0.514 (0.076)	0.369 (0.071)	0.988 (0.009)*	$3.7110e^{-01}$
RUSRF	0.906 (0.009)	0.84 (0.012)	0.583 (0.044)	0.708 (0.079)	0.865 (0.007)	$3.8670e^{-04***}$

Across these methods, RUSRF had the best AUC and F1 score.
Key: AUC, area under the curve; MMSE, mini-mental state examination; pMCI, progressive MCI; RUSRF, RUS random forest; SVM, support vector machine.
$^{*}p < 0.025$, $^{**}p < 0.0125$, and $^{***}p < 0.0025$.

minority class. In the present study, pMCI individuals were only 15.75% of the sample. Therefore, comparing the prediction accuracy based on the overall accuracy will lead to incorrect conclusions due to the adverse bias toward the majority class (sMCI). In other words, a classifier predicting sMCI, regardless of the input features, will result in 84.25% overall accuracy as there will only be 15.75% incorrect predictions (pMCI individuals predicted as sMCI; 0% accuracy in pMCI class). To this regard, we have excluded the overall accuracy as a measure to compare predictors used in this study; however, we used McNemar's χ^2 test to compare the sensitivity and the specificity. It is important to mention that most of the studies conducted to predict the progression of Alzheimer's disease dementia from MCI have not considered whether their samples represent the imbalanced proportions of the population (see Table 3). This can lead to inconsistent results when the studies are used in a clinical setting as the models have been trained and tested on a heightened representation of the pMCI population. Given that the population used in the present study is a closer representation of the MCI population, the results of the study can be considered transferable to the clinical use in early diagnosis.

McNemar's χ^2 test was used to evaluate the statistical significant difference between each predictor used in the study (RUSRF, SVM, regularized logistic regression, and random forest) against a random predictor. The random predictor was created by training each of the algorithms (RUSRF, SVM, regularized logistic regression, and random forest) with a randomly permuted set of features derived from the original training set. By doing this, the relationship between the features and the classes are completely disrupted and the predictor is trained to learn a random pattern. A statistically significant difference from a random predictor was only observed in the RUSRF predictor at $p < 0.01$, indicating that the novel algorithm was able to learn the classification patterns with a significant difference to a randomly trained classifier. Both regularized logistic regression algorithms showed differences at $p < 0.1$, indicating a trend. This may be due to the limited number of samples used in the testing set and the skewed proportions of sMCI and pMCI individuals. However, neither the random forest predictor nor the SVM predictor showed a significant difference from a random predictor, indicating their failure to learn the prediction relationships from the training set.

McNemar's χ^2 test and paired t-test were used to compare the performance of the RUSRF algorithm with other prediction algorithms. The AUC score can be considered a single metric for classifier performance when dealing with highly unequal sample sizes (Eskildsen et al., 2013). The RUSRF algorithm achieved the highest AUC score and was significantly different to both SVM and L2 logistic regression at $p < 0.0025$ and to random forest at $p < 0.0125$. Although the AUC of RUSRF was nominally higher than the AUC of L1 logistic regression, it was not significantly different. However, the standard deviation of the AUC in L1 logistic regression was considerably greater than that of RUSRF. This indicates that the performance of the L1 logistic regression is more sensitive to the noise presented in the input features (Frome et al., 2007; Fu et al., 2006), which can lead to inconsistent results in a clinical setting when used in early diagnosis. We also used the F1 score to compare the predictor performance as it considers both the precision and recall to construct the metric. The RUSRF predictor achieved the highest F1 score and was significantly different from SVM at $p < 0.0025$ and from L2 logistic regression at $p < 0.0125$. The F1 score of RUSRF was nominally higher compared with both L1 logistic regression and random forest but was not significant. However, in the case of random forest, this can be due to the high precision resulting from a high specificity even with a lower sensitivity value. It is important to mention that the F1 score can be influenced by skewed class proportions (Jeni et al., 2013) and needs to be considered when used as a performance measure. Both AUC and F1 scores were compared using paired t-tests with Bonferroni adjusted thresholds for inference.

Comparing the sensitivity and specificity between RUSRF and the other predictors was performed using the McNemar's χ^2 test. The sensitivity of the RUSRF predictor was significantly higher only when compared to the SVM predictor. This can be greatly due to the low number of pMCI individuals ($n = 13$) in the test set, as the statistic is calculated considering only the pMCI individuals. Furthermore, both L1 and L2 logistic regression predictors were nominally higher than the sensitivity of RUSRF. However, given the number of pMCI individuals, misclassification of even a single individual can cause the sensitivity metric to fluctuate substantially. Although the specificity of both the SVM and the random forest predictors were higher than RUSRF, their sensitivities were not acceptable for any clinical or research use. Among the RUSRF and logistic regression predictors, RUSRF achieved the highest specificity and was significantly different from L2 logistic regression at $p < 0.0125$ and from L1 logistic regression at $p < 0.025$. Although the SVM classifier is used widely in many applications, our analysis indicated that, at its current configuration with a radial basis kernel, it produced a very low sensitivity and is not suitable in identifying pMCI individuals. However, the SVM classifier can be trained with a number of kernel functions such as the polynomial kernel functions, sigmoid kernel functions, and linear kernel functions, and this would lead to increased sensitivity and overall classification performance.

Direct comparisons with other similar studies conducted to assess the predictive capability of amyloid imaging–based biomarkers cannot be performed due to a number of reasons (see summarized comparison in Table 3). Most studies (Brück et al., 2013; Hatashita and Yamasaki, 2013; Jack et al., 2010a,b; Wolk et al., 2009) have used [^{11}C]PiB PET as the imaging biomarker, as it was commonly used before the introduction of [^{18}F]-based radio ligands (except the authors of (Waragai et al., 2009), have used [^{11}C] BF-227 PET). Both Teipel et al. (2015) and Schreiber et al. (2015) have used [^{18}F]Florbetapir PET; however, Teipel et al. (2015) have used both [^{18}F]FDG PET and MRI biomarkers in combination with amyloid PET as the input features for the predictor. Furthermore, studies with similar observation periods (Brück et al., 2013; Ewers et al., 2012; Jack et al., 2010a,b; Waragai et al., 2009) have reported much higher progression rates. These increased rates may

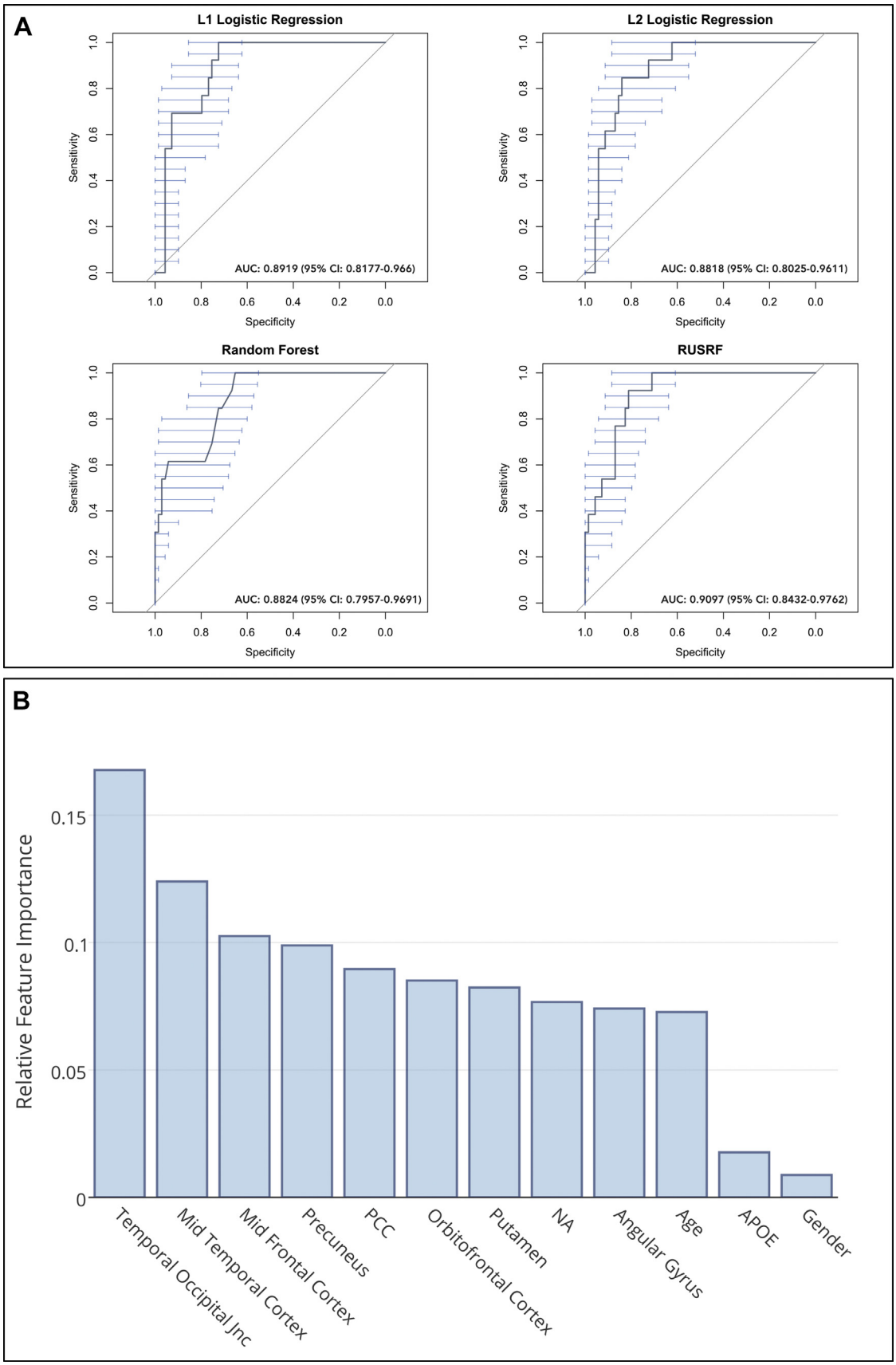

Fig. 3. (A) ROC curves for the 4 predictors with corresponding AUC values. The ROC curve was plotted for the median performing predictor from the cross-validation steps. The highest AUC was recorded with the curve for RUSRF. (B) Relative feature importance from the RUSRF predictor. The temporal occipital junction, the mid temporal cortex, the mid frontal cortex, and the precuneus show the highest importance for the prediction. Abbreviations: AUC, area under the curve; ROC, receiver operating characteristic; RUSRF, RUS random forest.

Table 3
Summary of several previous studies performed to predict progression to Alzheimer's disease dementia from MCI

Study	Data	Validation	Result (ACC—Accuracy, SEN—Sensitivity, SPE—Specificity)
Waragai et al., 2009	Total: 13, sMCI: 7, pMCI: 6 Follow-up: 27 mo [^{11}C]BF-227 PET	No validation	SEN: 100%, SPE: 71%
Hatashita and Yamasaki, 2013	Total: 68, sMCI: 38, pMCI: 30 Follow-up: 19.2 mo [^{11}C]PIB PET	No validation	SEN: 97%, SPE: 42%
Brück et al., 2013	Total: 29, sMCI: 12, pMCI: 17 Follow-up: 24 mo [^{11}C]PIB PET and [^{18}F]FDG PET	No validation	SEN: 88%, SPE: 71%
Wolk et al., 2009	Total: 26, sMCI: 21, pMCI: 5 Follow-up: 21.2 mo [^{11}C]PIB PET	No validation	SEN: 100%, SPE: 55%
Ewers et al., 2012	Total: 131, sMCI: 56, pMCI:75 Follow-up: 24 mo [^{11}C]PIB PET	No validation	SEN: 81%, SPE: 83%
Jack et al., 2010a,b	Total: 218, sMCI: 125, pMCI: 93 Follow-up: 24 mo [^{11}C]PIB PET	No validation	SEN: 50%, SPE: 81%
Schreiber et al., 2015	Total: 401, sMCI: 340, pMCI: 61 Follow-up: 19.2 mo [^{18}F]Florbetapir PET	No validation	SEN: 87%, SPE: 50%
Teipel et al., 2015	Total: 127, sMCI: 88, pMCI: 39 Follow-up: 17.3 mo [^{18}F]Florbetapir PET, [^{18}F]FDG PET and MRI	Cross validation	ACC: 72% SEN: 30%, SPE: 90% AUC: 0.70
Moradi et al., 2015	Total: 264, sMCI: 100, pMCI: 164 Follow-up: 0–36 mo MRI and cognition	Cross validation	ACC: 82% SEN: 87%, SPE: 74% AUC: 0.90
Misra et al., 2009	Total: 103, sMCI: 76, pMCI: 27 Follow-up: 0–36 mo MRI	Cross validation	ACC: 75%–80% AUC: 0.77
Davatzikos et al., 2011	Total: 239, sMCI: 170, pMCI: 69 Follow-up: 0–36 mo MRI and CSF	Cross validation	ACC: 62% AUC: 0.73
Ye et al., 2012	Total: 319, sMCI: 177, pMCI: 142 MRI, genetic and cognitive measures	Cross validation	AUC: 0.86
Zhang and Shen, 2012	Total: 88, sMCI: 50, pMCI: 38 Follow-up: 0–24 mo MRI, [18F]FDG PET and CSF	Cross validation	ACC: 78% SEN: 79%, SPE: 78% AUC: 0.77
Gaser et al., 2013	Total: 195, sMCI: 62, pMCI: 133 Follow-up: 0–36 mo MRI	Independent test set	AUC: 0.78
Cuingnet et al., 2011	Total: 210, sMCI: 134, pMCI: 76 Follow-up: 0–18 mo MRI	Independent test set	ACC: 67% SEN: 62%, SPE: 69%
Eskildsen et al., 2013	Total: 388, sMCI: 227, pMCI: 161 Follow-up: 0–48 mo MRI	Cross validation	AUC: 0.71[a]
Wolz et al., 2011	Total: 405, sMCI: 238, pMCI: 167 Follow-up: 0-48 moths MRI	Cross validation	ACC: 68% SEN: 67%, SPE: 69%
Chupin et al., 2009	Total: 210, sMCI: 134, pMCI: 76 Follow-up: 0–18 mo MRI	Independent test set	ACC: 64% SEN: 60%, SPE: 65%
Cho et al., 2012	Total: 203, sMCI: 131, pMCI: 72 Follow-up: 0–18 mo MRI	Independent test set	ACC: 71% SEN: 63%, SPE: 76%
Coupé et al., 2012	Total: 405, sMCI: 238, pMCI: 167 Follow-up: 0–48 mo MRI	Cross validation	ACC: 74% SEN: 73%, SPE: 74%
Westman et al., 2012	Total: 318, sMCI: 256, pMCI: 62 Follow-up: 0–12 mo MRI	Cross validation	ACC: 59% SEN: 74%, SPE: 56%
Querbes et al., 2009	Total: 122, sMCI: 50, pMCI: 72 Follow-up: 0–24 mo MRI	Cross validation	ACC: 73% SEN: 75%, SPE: 69%
Koikkalainen et al., 2011	Total: 369, sMCI: 215, pMCI: 154 Follow-up (mean): 18.1 mo MRI	Cross validation	ACC: 72% SEN: 77%, SPE: 71%
Retico et al., 2015	Total: 302, sMCI: 166, pMCI: 136 Follow-up: 0–24 mo MRI	Independent test set	ACC: 66% SEN: 70%, SPE: 62% AUC: 0.71

(*continued on next page*)

Table 3 (*continued*)

Study	Data	Validation	Result (ACC—Accuracy, SEN—Sensitivity, SPE—Specificity)
Doyle et al., 2014	Total: 226, sMCI: 164, pMCI: 62 Follow-up: 0–12 mo MRI	Independent test set (different cohort)	ACC: 82% SEN: 81%, SPE: 72% AUC: 0.81
Lebedev et al., 2014	Total: 139, sMCI: 16, pMCI: 123 Follow-up: 0–24 mo MRI	Independent test set	ACC: 82% SEN: 83%, SPE: 81% AUC: 0.83
This study	Total: 263, sMCI: 230, pMCI: 43 Follow-up: 24 mo [^{18}F]Florbetapir PET	Independent test set	ACC: 84% SEN: 71%, SPE: 87% AUC: 0.91

Key: CSF, cerebrospinal fluid; MRI, magnetic resonance imaging; pMCI, progressive MCI; sMCI, stable MCI.
[a] Follow-up period matched to present study.

be due to the low number of subjects included (Waragai et al., 2009) by specialized memory clinics (Brück et al., 2013) and the lower baseline Mini-Mental State Examination scores reported (Brück et al., 2013; Ewers et al., 2012; Jack et al., 2010a,b; Waragai et al., 2009).

Nevertheless, these studies except for Teipel et al. (2015) have not used any model evaluation technique, such as cross validation, to assess how their models will generalize to an independent data set, which is of paramount importance to be successfully used as an early diagnosis tool in a clinical environment. We obtained a higher overall accuracy, AUC, and sensitivity compared with the accuracy reported by Teipel et al. (2015). The low sensitivity values reported in Teipel et al. (2015) may be due the lower proportion of pMCI individuals included in the study, leading to a bias toward the majority sMCI group. The novel algorithm introduced in this study compensates for this "imbalanced class problem" by using data sampling techniques as mentioned earlier.

Compared with several previous studies using not only Aβ PET–based biomarkers, the algorithm introduced here seems promising with a recorded average AUC of 0.906 for an independent test set (accuracy: 0.84; sensitivity: 0.708; and specificity: 0.865). Often, studies use performance measures (AUC, accuracy, sensitivity, and specificity) based only on cross validation; however, this makes the prediction model biased toward the data set used as all the samples are used for training at some point in the analysis. The present study uses a completely independent testing set to evaluate the true generalization capability of the model presented. The independent testing set is not used in the feature extraction step or the model training steps. By evaluating the performance of the prediction model based on a completely independent testing set, we can retrieve the true expected performance of the model when used in a clinical environment as an early diagnostic tool. The generalization capability can be further improved by training the model on data acquired from multiple cohorts, as using a test set from the same cohort can add bias to the prediction model.

The most important characteristic of the present study is its ability to overcome the inherent imbalance of proportion between pMCI and sMCI individuals and performing comparably or even better than the previous studies. Studies conducted by Cho et al. (2012), Davatzikos et al. (2011), Misra et al. (2009) and Teipel et al. (2015) used cohorts with imbalanced proportions similar to the present study to train their prediction models, which resulted in lower sensitivity scores and lower AUC values.

Another important outcome from the present study is the relative predictive capability of the investigated brain regions (Fig. 3B) based on Aβ retention. Our study indicated that the 4 brain regions, the temporal occipital junction, mid temporal cortex, mid frontal cortex and precuneus to have the highest predictive ability. Other studies conducted with Aβ PET–based biomarkers also identified similar regions to have the highest difference between individuals progressing to Alzheimer's disease and sMCI individuals (Brück et al., 2013; Koivunen et al., 2011; Teipel et al., 2015). These regions are known to show Aβ deposition in relatively later stages of the course of Alzheimer's disease. This agrees with the idea that the regions with early deposition of Aβ such as the posterior cingulate cortex, already reaches a plateau in the MCI stage, adding little information to discriminate individuals progressing to Alzheimer's disease, whereas regions with later build-up are more relevant in discriminating stable and progressing MCI individuals. Interestingly, the feature selection technique identified subcortical structures such as the putamen and the nucleus accumbens to be used as features for classification. It is worth to emphasize that patients with AD have shown increased deposition of Aβ in these subcortical brain regions occurring later in the disease spectrum (Ishibashi et al., 2014). Furthermore, these regions receive projections from different prefrontal sub-regions, mainly from the medial prefrontal cortex (Ferry et al., 2000; Ongür and Price, 2000), and their associations with behavioral characteristics and cognitive performance are well known (de Jong et al., 2012; Diekhof et al., 2012). When evaluating the brain Aβ deposition, the consensus of the field of Alzheimer's disease is to use the average deposition from the brain regions characteristic of Alzheimer's disease. However, this method averages the Aβ deposition in brain regions which accumulate in different stages of Alzheimer's disease, and will result in a suboptimal discriminating capability compared with the method introduced in the article (see Supplementary Material for results of RUSRF performance with average brain Aβ deposition). Identifying the relative importance of Aβ deposition in various brain regions provides many benefits to research studies of anti-amyloid agents to evaluate the treatment's effect in preventing or slowing the progression to Alzheimer's disease.

It is important to mention several limitations of the present study. Any Alzheimer's disease study involving in vivo data suffers from the inherent uncertainty of the diagnosis (Eskildsen et al., 2013). This is because the diagnosis of Alzheimer's disease can only be confirmed with autopsy data. Therefore, any clinical diagnosis will always be "probable Alzheimer's disease." It is estimated that around 10% of this "probable Alzheimer's disease" diagnosis can be mislabeled "Lewy Body disease" or "Frontotemporal dementia" (Ranginwala et al., 2008). Thus, the patterns learnt by the predictor might not be specific to Alzheimer's disease; however, this has been partially addressed by using an Alzheimer's disease–specific biomarker such as Aβ PET compared with a general neurodegenerative biomarker such as MRI or glucose metabolism.

It is important to emphasize that the population included in this analysis represents a select group of amnestic MCI individuals motivated to participate in a dementia study, as such, for reasons related to the study inclusion criteria these individuals may not represent the general MCI population. Therefore, it would be highly desirable to replicate our findings in a population-based cohort.

Furthermore, the predictor presented here is biased to the [^{18}F] Florbetapir PET acquisition protocol followed by ADNI, as the training and testing was done using only the data acquired from the ADNI database. Hence, clinical implementations of the present study will still require additional reliability assessments by training the predictor using multiple acquisition protocols to increase the robustness of the prediction. Similarly, the feature extraction step included identifying brain regions with a scaled odds ratio above 1.5, based on the expert opinion. The rationale behind this step was to select the anatomically significant brain regions with an increase of odds for developing Alzheimer's disease of 50% for each 1 standard deviation increase of SUVR. However, this step can add bias toward the subjective opinion of the expert and can be avoided by selecting brain regions based on the opinion of multiple experts or by using an automated cluster–based technique. Furthermore, due to the computational cost of performing multiple voxel-wise logistic regression analyses, the feature selection technique employed a single voxel-wise logistic regression analysis including all the subjects (excluding the ones used as the final test dataset). However, the stability of the identified regions can be further increased by employing an undersampling technique in identifying the features.

In conclusion, we present a novel algorithm to predict the progression to Alzheimer's disease dementia within MCI individuals based on their Aβ PET measurements. The predictor presented here achieved a high rate of accuracy with an AUC of 0.906 for an independent test set. The novel algorithm overcomes the inherent imbalance of proportions between stable and pMCI seen in a population of MCI individuals, making it ideally suited for a clinical environment as an early diagnostic tool. Demonstration of such a tool using the data presented in the current article can be found at http://predictalz.tnlmcgill.ca/PredictAlz_Amy.

Disclosure statement

The authors have no actual or potential conflicts of interest.

Acknowledgements

Data collection and sharing for this project was funded by the 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; Eisai Inc; Elan Pharmaceuticals, Inc; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co, Inc; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research also provides 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 Disease Cooperative Study at the University of California, San Diego. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work was supported by the Canadian Institutes of Health Research (CIHR) [MOP-11-51-31], Canadian Consortium of Neurodegeneration and Aging, the Alan Tiffin Foundation, the Alzheimer's Association [NIRG-12-92090, NIRP-12-259245], the Fonds de Recherche du Québec–Santé (P-RN), and the Centre for Studies on Prevention of Alzheimer's Disease (StoP-AD Centre).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.neurobiolaging.2017.06.027.

References

Ad-Dab'bagh, Y., Einarson, D., Lyttelton, O., Muehlboeck, J.S., Mok, K., Ivanov, O., Vincent, R.D., Lepage, C., Lerch, J., Fombonne, E., Evans, A.C., 2006. The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. In: Corbetta, M. (Ed.), Proceedings of the 12th Annual Meeting of the Human Brain Mapping Organization. Neuroimage, Florence, Italy.

Apostolova, L.G., Hwang, K.S., Kohannim, O., Avila, D., Elashoff, D., Jack, C.R., Shaw, L., Trojanowski, J.Q., Weiner, M.W., Thompson, P.M., 2014. ApoE4 effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer's disease. Neuroimage Clin. 4, 461–472.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82, 239–259.

Breiman, L., 2001. Random forests. Mach. Learn 45, 5–32.

Brück, A., Virta, J.R., Koivunen, J., Koikkalainen, J., Scheinin, N.M., Helenius, H., Någren, K., Helin, S., Parkkola, R., Viitanen, M., Rinne, J.O., 2013. [11C]PIB, [18F] FDG and MR imaging in patients with mild cognitive impairment. Eur. J. Nucl. Med. Mol. Imaging 40, 1567–1572.

Buerger, K., Ewers, M., Pirttilä, T., Zinkowski, R., Alafuzoff, I., Teipel, S.J., DeBernardis, J., Kerkman, D., McCulloch, C., Soininen, H., Hampel, H., 2006. CSF phosphorylated tau protein correlates with neocortical neurofibrillary pathology in Alzheimer's disease. Brain 129, 3035–3041.

Cho, Y., Seong, J.-K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage 59, 2217–2230.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19, 579–587.

Collins, D.L., Evans, A.C., 1997. Animal: validation and applications of nonlinear registration-based segmentation. Int. J. Pattern Recognit. Artif. Intell. 11, 1271–1294.

Collins, D.L., Holmes, C.J., Peters, T.M., Evans, A.C., 1995. Automatic 3-D model-based neuroanatomical segmentation. Hum. Brain Mapp. 3, 190–208.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59, 3736–3747.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56, 766–781.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32, 2322.e19–2322.e27.

de Jong, L.W., Wang, Y., White, L.R., Yu, B., van Buchem, M.A., Launer, L.J., 2012. Ventral striatal volume is associated with cognitive decline in older people: a population based MR-study. Neurobiol. Aging 33, 424.e1–424.e10.

Diekhof, E.K., Kaps, L., Falkai, P., Gruber, O., 2012. The role of the human ventral striatum and the medial orbitofrontal cortex in the representation of reward magnitude? An activation likelihood estimation meta-analysis of neuroimaging studies of passive reward expectancy and outcome processing. Neuropsychologia 50, 1252–1266.

Doyle, O.M., Westman, E., Marquand, A.F., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Williams, S.C.R., Simmons, A., 2014. Predicting progression of Alzheimer's disease using ordinal regression. PLoS One 9, e105542.

Eskildsen, S.F., Coupé, P., García-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.

Ewers, M., Insel, P., Jagust, W.J., Shaw, L., Trojanowski, J.Q., Aisen, P., Petersen, R.C., Schuff, N., Weiner, M.W., 2012. CSF biomarker and PIB-PET-derived beta-amyloid signature predicts metabolic, gray matter, and cognitive changes in nondemented subjects. Cereb. Cortex 22, 1993–2004.

Fagan, A.M., Mintun, M.A., Mach, R.H., Lee, S.-Y., Dence, C.S., Shah, A.R., LaRossa, G.N., Spinner, M.L., Klunk, W.E., Mathis, C.A., DeKosky, S.T., Morris, J.C.,

Holtzman, D.M., 2006. Inverse relation between in vivo amyloid imaging load and cerebrospinal fluid Abeta42 in humans. Ann. Neurol. 59, 512–519.
Ferry, A.T., Ongür, D., An, X., Price, J.L., 2000. Prefrontal cortical projections to the striatum in macaque monkeys: evidence for an organization related to prefrontal networks. J. Comp. Neurol. 425, 447–470.
Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. Neuroimage 47, S102.
Frome A., Sha F., Singer Y., Malik J., Sha F., Malik J., Learning globally-consistent local distance functions for shape-based image retrieval and classification, In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on IEEE, pp. 1–8. http://dx.doi.org/10.1109/ICCV.2007.4408839.
Fu, H., Ng, M.K., Nikolova, M., Barlow, J.L., 2006. Efficient minimization methods of mixed l2-l1 and l1-l1 norms for image restoration. SIAM J. Sci. Comput. 27, 1881–1902.
Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PLoS One 8, e67346.
Gelosa, G., Brooks, D., 2012. The prognostic value of amyloid imaging. Eur. J. Nucl. Med. Mol. Imaging 39, 1207–1219.
Hall, A., Mattila, J., Koikkalainen, J., Lötjonen, J., Wolz, R., Scheltens, P., Frisoni, G., Tsolaki, M., Nobili, F., Freund-levi, Y., Minthon, L., Frölich, L., Hampel, H., Visser, P.J., Soininen, H., 2015. Predicting progression from cognitive impairment to Alzheimer's disease with the Disease State Index. Curr. Alzheimer Res. 12, 69–79.
Hatashita, S., Yamasaki, H., 2013. Diagnosed mild cognitive impairment due to Alzheimer's disease with PET biomarkers of beta amyloid and neuronal dysfunction. PLoS One 8, e66877.
Holland, D., McEvoy, L.K., Desikan, R.S., Dale, A.M., 2012. Enrichment and stratification for predementia Alzheimer disease clinical trials. PLoS One 7, e47739.
Huttunen, H., Manninen, T., Kauppi, J.-P., Tohka, J., 2012. Mind reading with regularized multinomial logistic regression. Mach. Vis. Appl. 24, 1311–1325.
Ishibashi, K., Ishiwata, K., Toyohara, J., Murayama, S., Ishii, K., 2014. Regional analysis of striatal and cortical amyloid deposition in patients with Alzheimer's disease. Eur. J. Neurosci. 40, 2701–2706.
Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, V.S., Donohue, M.C., Trojanowski, J.Q., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12, 207–216.
Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010a. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9, 119–128.
Jack, C.R., Wiste, H.J., Vemuri, P., Weigand, S.D., Senjem, M.L., Zeng, G., Bernstein, M.A., Gunter, J.L., Pankratz, V.S., Aisen, P.S., Weiner, M.W., Petersen, R.C., Shaw, L.M., Trojanowski, J.Q., Knopman, D.S., 2010b. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. Brain 133, 3336–3348.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.
Jeni, L.A., Cohn, J.F., De La Torre, F., 2013. Facing imbalanced data–recommendations for the use of performance metrics. 2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact 245–251.
Koikkalainen, J., Lötjönen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., 2011. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. Neuroimage 56, 1134–1144.
Koivunen, J., Scheinin, N., Virta, J.R., Aalto, S., Vahlberg, T., Nagren, K., Helin, S., Parkkola, R., Viitanen, M., Rinne, J.O., Någren, K., Helin, S., Parkkola, R., Viitanen, M., Rinne, J.O., 2011. Amyloid PET imaging in patients with mild cognitive impairment: a 2-year follow-up study. Neurology 76, 1085–1090.
Lebedev, A.V., Westman, E., Van Westen, G.J.P., Kramberger, M.G., Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., Vellas, B., Lovestone, S., Simmons, A., 2014. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuroimage Clin. 6, 115–125.
Markesbery, W.R., 2010. Neuropathologic alterations in mild cognitive impairment: a review. J. Alzheimer's Dis. 19, 221–228.
Mathotaarachchi, S., Wang, S., Shin, M., Pascoal, T.A., Benedet, A.L., Kang, M.S., Beaudry, T., Fonov, V.S., Gauthier, S., Labbe, A., Rosa-Neto, P., 2016. VoxelStats: a MATLAB package for multi-modal voxel-wise brain image analysis. Front. Neuroinform 10, 20.
Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J., 1995. A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). Neuroimage 2, 89–101.
McEvoy, L.K., Brewer, J.B., 2010. Quantitative structural MRI for early detection of Alzheimer's disease. Expert Rev. Neurother. 10, 1675–1688.
Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 44, 1415–1422.
Mitchell, A.J., Shiri-Feshki, M., 2009. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. Acta Psychiatr. Scand. 119, 252–265.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Morris, J.C., Roe, C.M., Grant, E.A., Head, D., Storandt, M., Goate, A.M., Fagan, A.M., Holtzman, D.M., Mintun, M.A., 2009. Pittsburgh compound B imaging and prediction of progression from cognitive normality to symptomatic Alzheimer disease. Arch. Neurol. 66, 1469–1475.
Mosconi, L., Brys, M., Glodzik-Sobanska, L., De Santi, S., Rusinek, H., de Leon, M.J., 2007. Early detection of Alzheimer's disease using neuroimaging. Exp. Gerontol. 42, 129–138.
Ongür, D., Price, J.L., 2000. The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. Cereb. Cortex 10, 206–219.
Petersen, R.C. (Ed.), 2003. Mild Cognitive Impairment: Aging to Alzheimer's Disease. Oxford University Press, Oxford.
Pouryamout, L., Dams, J., Wasem, J., Dodel, R., Neumann, A., 2012. Economic evaluation of treatment options in patients with Alzheimer's disease: a systematic review of cost-effectiveness analyses. Drugs 72, 789–802.
Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132, 2036–2047.
Ranginwala, N.A., Hynan, L.S., Weiner, M.F., White, C.L., 2008. Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years. Am. J. Geriatr. Psychiatry 16, 384–388.
Retico, A., Bosco, P., Cerello, P., Fiorina, E., Chincarini, A., Fantacci, M.E., 2015. Predictive models based on support vector machines: whole-brain versus regional analysis of structural MRI in the Alzheimer's disease. J. Neuroimaging 25, 552–563.
Robbins, S., Evans, A.C., Collins, D.L., Whitesides, S., 2004. Tuning and comparing spatial normalization methods. Med. Image Anal. 8, 311–323.
Schreiber, S., Landau, S.M., Fero, A., Schreiber, F., Jagust, W.J., 2015. Comparison of visual and quantitative Florbetapir F 18 positron emission tomography analysis in predicting mild cognitive impairment outcomes. JAMA Neurol. 72, 1183–1190.
Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2008. RUSBoost: Improving classification performance when training data is skewed. In: 2008 19th International Conference on Pattern Recognition. IEEE, pp. 1–4.
Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2010. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man. Cybern. Part A Syst. Humans 40, 185–197.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.
Teipel, S.J., Kurth, J., Krause, B., Grothe, M.J., 2015. The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - beyond classical regression. Neuroimage Clin. 8, 583–593.
Trzepacz, P.T., Yu, P., Sun, J., Schuh, K., Case, M., Witte, M.M., Hochstetler, H., Hake, A., 2014. Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. Neurobiol. Aging 35, 143–151.
Waragai, M., Okamura, N., Furukawa, K., Tashiro, M., Furumoto, S., Funaki, Y., Kato, M., Iwata, R., Yanai, K., Kudo, Y., Arai, H., 2009. Comparison study of amyloid PET and voxel-based morphometry analysis in mild cognitive impairment and Alzheimer's disease. J. Neurol. Sci. 285, 100–108.
Westman, E., Muehlboeck, J.-S., Simmons, A., 2012. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 62, 229–238.
Wolk, D.A., Price, J.C., Saxton, J.A., Snitz, B.E., James, J.A., Lopez, O.L., Aizenstein, H.J., Cohen, A.D., Weissfeld, L.A., Mathis, C.A., Klunk, W.E., De-Kosky, S.T., DeKoskym, S.T., 2009. Amyloid imaging in mild cognitive impairment subtypes. Ann. Neurol. 65, 557–568.
Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS One 6, e25446.
Wolz, R., Schwarz, A.J., Gray, K.R., Yu, P., Hill, D.L.G. Alzheimer's Disease Neuroimaging Initiative, 2016. Enrichment of clinical trials in MCI due to AD using markers of amyloid and neurodegeneration. Neurology 87, 1235–1241.
Yang, X., Tan, M.Z., Qiu, A., 2012. CSF and brain structural imaging markers of the Alzheimer's pathological cascade. PLoS One 7, e47406.
Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A., 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 12, 46.
Yen, S.J., Lee, Y.S., 2009. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 36, 5718–5727.
Young, J., Modat, M., Cardoso, M.J., Ashburner, J., Ourselin, S., 2012. Classification of Alzheimer's disease patients and controls with Gaussian processes. 2012 9th IEEE Int. Symp. Biomed. Imaging 1523–1526.
Zhang, D., Shen, D., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7, e33182.
Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imaging 21, 1280–1291.