Journal of Alzheimer's Disease xx (2021) x–xx
DOI 10.3233/JAD-201398
IOS Press

A Role for Prior Knowledge in Statistical Classification of the Transition from Mild Cognitive Impairment to Alzheimer's Disease

Zihuan Liu[a], Tapabrata Maiti[a], Andrew R. Bender[b,*] and for the Alzheimer's Disease Neuroimaging Initiative[1]
[a]*Department of Statistics, Michigan State University, East Lansing, MI, USA*
[b]*Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, USA*

Handling Associate Editor: Ali Ezzati

Accepted 2 August 2021
Pre-press 25 August 2021

Abstract.
Background: The transition from mild cognitive impairment (MCI) to dementia is of great interest to clinical research on Alzheimer's disease and related dementias. This phenomenon also serves as a valuable data source for quantitative methodological researchers developing new approaches for classification. However, the growth of machine learning (ML) approaches for classification may falsely lead many clinical researchers to underestimate the value of logistic regression (LR), which often demonstrates classification accuracy equivalent or superior to other ML methods. Further, when faced with many potential features that could be used for classifying the transition, clinical researchers are often unaware of the relative value of different approaches for variable selection.
Objective: The present study sought to compare different methods for statistical classification and for automated and theoretically guided feature selection techniques in the context of predicting conversion from MCI to dementia.
Methods: We used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to evaluate different influences of automated feature preselection on LR and support vector machine (SVM) classification methods, in classifying conversion from MCI to dementia.
Results: The present findings demonstrate how similar performance can be achieved using user-guided, clinically informed pre-selection versus algorithmic feature selection techniques.
Conclusion: These results show that although SVM and other ML techniques are capable of relatively accurate classification, similar or higher accuracy can often be achieved by LR, mitigating SVM's necessity or value for many clinical researchers.

Keywords: Alzheimer's disease, classification, machine learning, mild cognitive impairment, support vector machine, variable selection

*Correspondence to: Andrew R. Bender, Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, USA. E-mail: arbender@msu.edu.

[1]Data used in preparation of this article were 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

INTRODUCTION

Alzheimer's disease (AD) is a progressive, age-related, neurodegenerative disease and the most common cause of dementia [1–3]. Behaviorally, AD is commonly preceded by mild cognitive impairment (MCI), a syndrome characterized by declines in memory and other cognitive domains that exceed cognitive decrements associated with normal aging [2, 4]. However, the prodromal symptoms of MCI are not prognostically deterministic: individuals with MCI tend to progress to diagnoses of probable AD at a rate of 8%–15% per year, and many conversions are detectable within 3 years of initial presentation [5–7]. Research efforts to provide new insights into the incidence of MCI-to-AD conversion have focused largely on clinically or biologically relevant features (i.e., neuroimaging markers, clinical exam data, neuropsychological test scores) and on different methods for statistical classification [8].

For clinical researchers, however, there may be a tendency to conflate more sophisticated, novel analytic approaches and the value of multimodal information from neuroimaging and clinical assessment. Moreover, whereas statisticians may inherently understand the comparability of different quantitative approaches, the novelty of both big data and data-driven approaches for studying MCI-to-AD conversion may lead clinical researchers to assume that such data-driven methods are inherently superior to more theoretically grounded approaches. Thus, the value of using extant findings and domain expertise to help guide and constrain the application of newer data-driven approaches capable of capitalizing on emergent big data may be a particularly important consideration for clinical researchers.

Statistical classification in clinical research has traditionally utilized binary logistic regression (LR). However, key attributes of modern clinical and neuroimaging data, including high dimensionality and the presence of ground truth estimates of pathology and diagnosis provide new opportunities for quantitative research. This has led to a substantial expansion in the use of data from the Alzheimer's Disease Neuroimaging Initiative (ADNI; http://adni.loni.usc.edu) for quantitative research and methodological development, particularly by researchers utilizing and developing prediction and classification methods in machine learning (ML). Besides LR, support vector machine (SVM) has quickly become the most common type of ML classifier for diagnostic prediction and classification with ADNI data. In general, LR works well when the data is linearly separable, and the number of data is greater than the number of features. Moreover, SVM and LR have similar misclassification rates (MCRs) when used to diagnose malignant tumors from imaging data [9, 10].

Indeed, before the rapid expansion of ML research and applied work over the past decade, many clinical researchers and those outside of engineering and mathematically intensive disciplines had little exposure to classification approaches other than LR. Despite its growing popularity, the relative benefits of SVM or other forms of ML [11, 12] over LR for such classification are not always apparent. Although this may be of little surprise to statisticians and quantitative researchers, such perspectives are often lost on clinical researchers, whose implicit beliefs in the superiority of ML is driven by the volume of publications, rather than through training or empirical demonstration.

Most efforts to develop new classification methods for prediction of MCI-to-AD conversion are well suited to integrate measures from multiple sources such as demographics, clinical rating scores, neuropsychological testing, neuroimaging, genetic markers, etc. However, identifying which combination of features most accurately classifies conversion from MCI to AD is a key challenge for ADNI, and may vary by method. The L_1 norm regularization method (i.e., L_1) is a highly used feature selection technique for LR and SVM. L_1 is popular for addressing circumstances in which the number of features is quite large or even larger than the sample size. Despite some risk of abusing the statistical terminology, the problem is often generically referred to as the "small n, large p" or high dimensional problem. The L_1 technique has dual impacts, namely the algorithm can (i) optimize a higher number of parameters in comparison to sample size, and (ii) reduce the effective number of parameters (i.e., performing variable selection). This powerful technique has been implemented in ADNI data with LR [13]. Furthermore, L_1 and other algorithmic feature selection methods used in ML suffer from one key limitation: they are agnostic to theoretical considerations, and as such, they cannot interpret why selected features are meaningful and important to the model. When sampling from a large pool of features, the algorithmic approaches fail to consider prior knowledge of features and their associations with the relevant systems in variable selection. Therefore, domain expertise and prior knowledge may afford additive or differential value for choosing features and interpreting model

results over algorithmic feature selection methods alone.

However, most real-world problems occur in the context of additional information about each potential feature and its conceptual relationship with the phenomenon being classified. Other than using L_1 feature selection, manually trimming the list of potential predictor variables can also protect against over-fitting, and also offers potential insight into why selected features are important to the model. When guided by prior knowledge, user-guided or 'manual' feature selection may be a valuable additional step to help minimize potentially spurious effects. This perspective is frequently lost on applied researchers, as most commonly used variable selection algorithms are context-free—that is, they only look at relationships within the data set, and cannot factor in the wider meanings of variables. Furthermore, this also means that automated algorithms may identify relationships among a large number of predictor variables that are spurious and are unlikely to generalize outside the data set. Although there are a vast number of potential neuroimaging features in ADNI data, the present study focused only on regional brain volumes segmented from structural magnetic resonance imaging (MRI) data, the most common neuroimaging datatype for classifying MCI-to-dementia conversion. In contrast to prior studies that used a limited set of volumetric brain features, the present study utilized data generated by modern multi-atlas segmentation methods and analyses included up to 259 features—anatomically specific gray and white matter volumes. However, the large pool of extant findings from studies evaluating regional brain MRI volumetry in prediction and classification of MCI-to-dementia conversion using both limited and expansive feature sets also provides a valuable set of priors for relevant brain regions [14–19]. Thus, applied researchers are often left with the conundrum of more confirmatory approaches that use few regions in classification or more exploratory methods in which prior findings have little value.

The present study addressed two questions regarding commonly used classification approaches for predicting MCI-to-dementia conversion in multimodal data from ADNI. First, we compared performance accuracy of binary LR with SVM in classifying MCI-to-dementia conversion. Second, we asked if applying prior knowledge in feature selection outperforms algorithmic variable selection alone. We hypothesized that 1) LR would perform comparably to SVM, and 2) user-guided variable selection would outperform algorithmic variable selection alone. This work is intended to demonstrate to clinical researchers the benefit of using ML in an informed fashion, rather than as a 'black box' that obscures clear interpretation. Moreover, we wish to emphasize that this study is not meant to highlight a novel innovation in quantitative methods, but rather to provide an important example to applied researchers regarding the comparable value of ML methods and importance of domain expertise in classification with ADNI data.

MATERIALS AND METHODS

Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu). The 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 AD. For up-to-date information, see http://www.adni-info.org.

Determination of sensitive and specific markers of preclinical AD and MCI is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as reduce the time and cost of clinical trials. Data in the present study came from all sites across the U.S and Canada. All ADNI study participants included in the present analyses were between 55 and 90 years old, spoke English or Spanish as their native language, and had a study partner who provided an independent assessment of functioning.

This study used a subset of the 819 participants from ADNI-1 diagnosed with MCI at baseline and for whom the data from demographic, clinical cognitive assessments, *APOE4* genotyping, and MRI measurements were also available. To evaluate differences in classification performance due to participant inclusion and drop out, we subdivided the sample into two overlapping groups. After applying other criteria for inclusion, Group One included all patients whose follow-up period was at least 36 months ($n = 265$); Group Two consisted of all patients with follow-up assessments at 24 months ($n = 308$). Although the ADNI study protocol includes additional follow-up visits at 6-month intervals, the present study only evaluated baseline data for features (i.e., clinical,

Table 1
Sample Sizes by Timing and Diagnosis: Group One and Two

Group	Time	# MCI-S ($y=0$)	# MCI-C ($y=1$)	# Total patients
One	36 months	101	164	265
Two	24 months	122	186	308

Table 1 shows the number of MCI-C, MCI-S, and total subjects in Group One and Two. The number of MCI-C patients is higher than MCI-S patients in both groups.

neuropsychological, brain volumetric) in classification analyses. In addition, identification of stable versus converting clinical outcomes only considered longer-term outcomes based on assessments at 2 and 3 years after baseline. The final samples included 265 and 308 study participants in Groups One and Two, respectively, who met criteria for inclusion. Both Groups included participants who were stable in their diagnosis (MCI-S) and those who converted to a diagnosis of dementia over the 2 or 3 years (MCI-C). Table 1 shows the participant characteristics. Diagnostic criteria for MCI included a Mini-Mental State Examination (MMSE) score at baseline between 24 and 30, a Clinical Dementia Rating (CDR) score of 0.5, and a subjective memory complaint, in addition to objective memory loss measured by education-adjusted scores on the Logical Memory II subscale of the Wechsler Memory Scale, generally preserved activities of daily living and no dementia. The diagnostic criteria for dementia were an MMSE score between 20 and 26, and a CDR score between 0.5 and 1.0. The clinical status of each participant diagnosed with MCI was re-assessed at each follow-up visit and updated to reflect one of several outcomes (e.g., MCI or dementia subtypes). The MCI-C and MCI-S group designations were based on this follow-up clinical diagnosis and marked as either 1 for MCI-C or 0 for MCI-S in classification study.

Data used in classification

Evaluation of extant reports of common predictors of conversion from MCI to dementia focused on dimensions of neuropsychological test performance, clinical assessment, genetic data, and regional brain volumes. In the present study, we first divided these variables into two sets of features, with all non-brain volumetric variables in one set and all variables representing regional brain volumes in a second set. In addition, we created a third set of features from the volumetry feature set that only included 26 of the 259 brain volumes. Henceforth, we refer to models that only include one of these three feature sets as 'single-modality,' whereas models that combine brain and non-brain feature sets are referred to as 'multi-modal.'

Clinical cognitive assessment and genetic data

We considered a total of 19 clinical features as potential predictors of MCI-to-AD progression in our classification analyses. These included the following assessment scores: the MMSE, CDR-Sum of Boxes, Alzheimer's Disease Assessment Scale-cognitive sub-scale (ADAS-cog), Functional Activities Questionnaire (FAQ) measures of activities of daily living, Trail Making Test-B (TRABSCOR), the immediate and delayed recall components of the Rey Auditory Verbal Learning Test (RAVLT), the Digit-Symbol Coding test (DIGT), and the Digit Symbol Substitution Test from the Preclinical Alzheimer Cognitive Composite (mPACCdigit). We also considered genotype for carriers of the epsilon-4 allele of the apolipoprotein E (*APOE*) gene [8] as a genetic predictor in this study. Table 2 summarizes all 19 clinical, demographic, and genetic features used in this study. Preliminary comparison of six clinical and genetic predictors by MCI-C and MCI-S subgroups showed five of them (APOE4, ADAS4, CDR, MMSE, and RAVLT.learning) significantly differ between the groups, whereas one (SEX) does not. Figures 1 and 2 illustrate the distribution of these predictors for both groups. Overall, in comparison to MCI-S participants, those in the MCI-C group were more cognitively and functionally impaired at baseline, exhibited greater verbal memory impairments, and included a greater proportion of *APOE4* carriers.

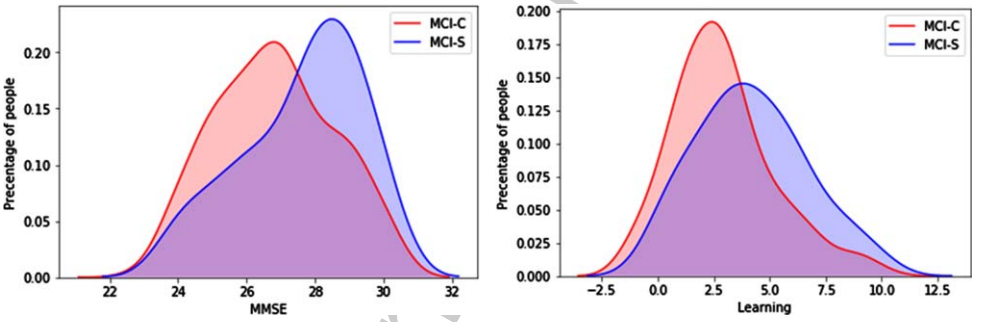
MRI data

Structural MRI data were collected according to the ADNI acquisition protocol using T1-weighted scans (GradWarp, B1 Correction, N3, Scaled) [20]. These data included baseline structural MRI scans of 840 ADNI participants, including 230 diagnosed as cognitively normal, 200 with diagnoses of dementia, and 410 diagnosed with MCI. Processing for region-of-interest (ROI)-based volumetric data used in the present study included brain extraction [21] and a multi-atlas, consensus-based label fusion scheme for anatomical parcellation [22] to generate template-based ROIs deformed to individual subject space. MRI scans were automatically segmented into

Table 2
Clinical Features and Cognitive Assessment Score of Group One

Characteristics	MCI-S	MCI-C	Test statistic	P
Age (y)	74.34 ± 7.78	74.84 ± 6.83	–0.528	>0.5[a]
Education (y)	15.57 ± 2.94	15.73 ± 2.91	–0.527	>0.5[b]
Sex, % female	33.67%	34.14%	0	1[b]
APOE4 carriers %	34.65%	62.19%	17.900	<0.001[a]
CDRSB	1.23 ± 0.61	1.72 ± 0.92	–5.237	<0.001[a]
MMSE score	27.61 ± 1.74	26.82 ± 1.71	3.645	<0.001[a]
ADAS11	8.89 ± 3.79	12.29 ± 4.16	–6.823	<0.001[a]
ADAS13	14.48 ± 5.50	20.01 ± 5.79	–7.795	<0.001[a]
ADASQ4	4.76 ± 2.19	6.77 ± 2.21	–7.339	<0.001[a]
RAVLT.immediate	36.21 ± 10.10	29.10 ± 7.98	6.021	<0.001[a]
RAVLT.learning	4.19 ± 2.74	2.91 ± 2.26	4.231	<0.001[a]
RAVLT.forgetting	4.31 ± 2.59	4.47 ± 2.15	–1.501	<0.135[a]
RAVLT.perc.forgetting	51.55 ± 31.04	72.85 ± 30.45	–5.464	<0.001[a]
LEDLTOTAL	4.96 ± 2.36	3.41 ± 2.66	4.931	<0.001[a]
DIGTSCOR	40.75 ± 11.09	36.62 ± 10.96	2.882	<0.005[a]
TRABSCOR	109.43 ± 62.94	132.09 ± 71.36	–2.704	<0.007[a]
FAQ	1.50 ± 2.99	4.96 ± 4.79	–7.243	<0.001[a]
mPACCdigit	–5.38 ± 2.96	–8.06 ± 2.96	7.174	<0.001[a]
mPACCtrailsB	–5.47 ± 3.06	–8.22 ± 2.98	7.174	<0.001[a]

Table only for Group One where has 265 patients and 36 months follow-up time. Values are shown as mean ± standard deviation or percentage. Test statistics and *p*-values for differences between MCI-S and MCI-C are based on (a) *t*-test or (b) chi- square test. MCI-S, non-progressive MCI; MCI-P, progressive MCI; *APOE*, apolipoprotein E; MMSE, Mini-Mental State Examination; RAVLT, The Rey Auditory Verbal Learning Test (immediate: sum of 5 trails; learning: trial 5-trial 1; Forgetting: trial 5-delayed; perc.forgetting: Percent forgetting); DIGT, The Digit- Symbol Coding test; TRAB, Trail Making tests; CDRSB, Clinical Dementia Rating Scaled Response; FAQ, Activities of Daily living Score; ADAS, Alzheimer's Disease Assessment Scale–Cognitive sub-scale; mPACCdigit, the Digit Symbol Substitution Test from the Preclinical Alzheimer Cognitive Composite.

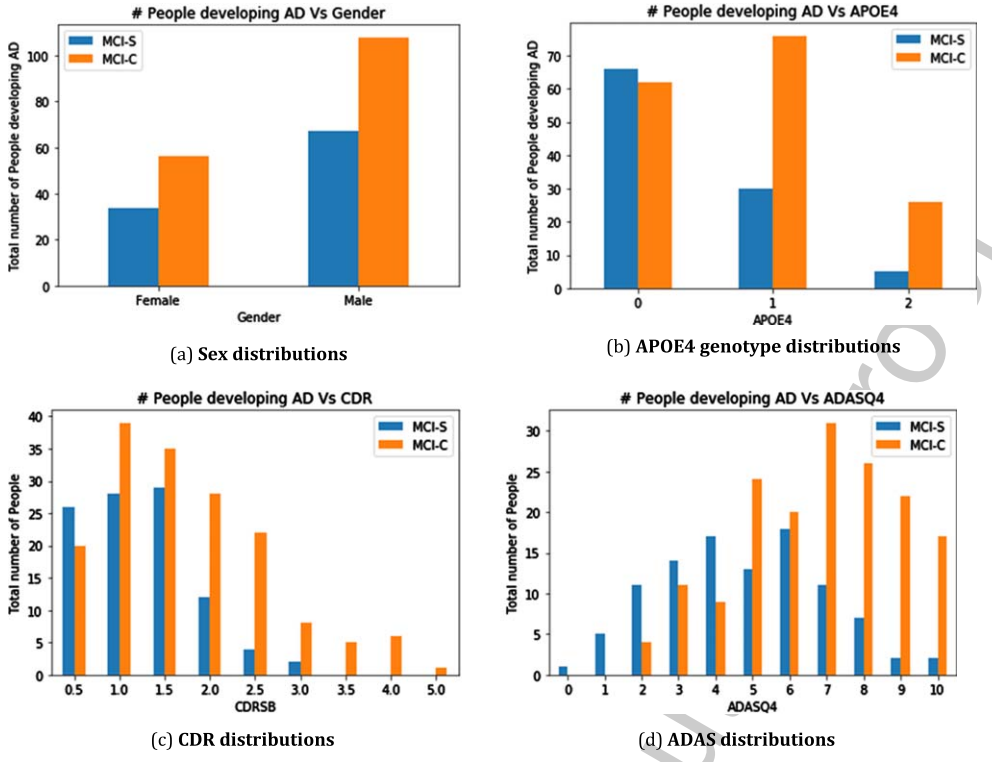


(a) **MMSE scores in MCI-C and MCI-S groups** (b) **Learning in MCI-C and MCI-S groups**

Fig. 1. Comparison of distributions for baseline predictor variables between MCI-S and MCI-C groups. (a) The mean MMSE score in MCI-S is higher than in MCI-C. (b) Mean Learning scores of MCI-C and MCI-S groups are 2.5 and 5.

145 anatomic ROIs spanning the entire brain. An additional 114 derived ROIs were calculated by combining single ROIs within a tree hierarchy, to obtain volumetric measurements from larger structures [20]. In total, 259 ROIs were measured and used as potential predictors of MCI-to-dementia progression in this study.

One of the goals of this study is to investigate if manually selecting predictors improves a model's performance. Based on the extant literature [23], we manually selected 26 out of 259 features as theoretically significant predictors of MCI to dementia progression (Table 3) [14–19]. While many brain regions have been reported as showing some relationship to MCI-to-dementia progression, prior reports and reviews clearly implicate hippocampal and entorhinal cortical volumes as markers of such conversion. In addition, we manually selected additional



(a) **Sex distributions** (b) **APOE4 genotype distributions**

(c) **CDR distributions** (d) **ADAS distributions**

Fig. 2. Comparisons between MCI-S and MCI-C groups on baseline predictor variables. The y-axis of panels (a) through (d) represents the number of participants developing AD. Blue and red bars represent non-converters and converters, respectively. Panel (a) shows a greater number of converters than non-converters for both men and women. Panel (b) shows more than half of MCI-C subjects are *APOE4* carriers and approximately 70% MCI-S subjects are non-*APOE4* carriers. Panel (c) shows MCI-S subjects have the relatively lower CDR score and MCI-C subjects have higher CDR score. The number of people in MCI-C group has a downward trend as CDR score increases. Panel (d) shows MCI-C subjects have the relatively higher ADASQ4 score. The average of ASADQ4 score of MCI-S and MCI-C subjects are approximately 5 and 8, respectively.

regions based on their common occurrence across reports, including cingulate gyrus, precuneus, amygdala, inferior frontal gyrus, superior parietal lobule, and lobar white matter volumes.

Method and algorithm

In the following section, we utilize binary LR and SVM classification techniques to investigate which approach yields superior discrimination accuracy in the context of ADNI data. Prior comparisons of logistic regression and SVM have reported that SVM requires fewer variables than logistic regression to achieve an equivalent level of MCR [10, 24]. These also report SVM performs better than LR with microarray expression data [10]. Furthermore, SVMs have a nice dual form, giving sparse solutions when using the kernel trick. In addition, both methods involve minimizing some cost associated with the misclassification based on likelihood ratio for a probabilistic model. Therefore, LR and SVM share common roots in statistical pattern recognition, which we utilize in the comparison of their performance on multi-modal ADNI data.

Logistic regression

LR is the most commonly used machine learning approach for binary classification. In the past decade this has been applied to task of MCI-to-dementia conversion [13, 25, 26]. In the present study, we consider a supervised learning task where we are given M training examples $D = (x_i, y_i)$, $i = 1, \ldots$ M. Here each $x_i\ \varepsilon R^N$ is N dimensional feature vectors, and $y_i\ \varepsilon\ 0,1$ is a class label. The goal of LR is to model the probability p of a random variable y being 1 or 0 given the experimental data x. The logistic regression model is defined as follows:

$$logit\ p = \log \frac{p}{1-p} \qquad (1)$$

Table 3
Pre-selected MRI features of Group One

Characteristics	MCI-S	MCI-C	Test statistic	p
HippoR	3684 ± 438	3366 ± 437	5.735	< 0.001
HippoL	3414 ± 418	3105 ± 388	5.994	< 0.001
flWMR	96720 ± 6218	96976 ± 5585	–0.338	0.73
flWML	93671 ± 5836	94238 ± 5160	–0.802	0.42
plWMR	47197 ± 3415	47141 ± 3098	0.135	0.89
plWML	50149 ± 3714	50038 ± 4367	0.242	0.81
tlWMR	56076 ± 3252	55934 ± 2931	0.359	0.72
tlWML	55412 ± 3396	55468 ± 3023	–0.136	0.89
ACgCR	3167 ± 756	3128 ± 641	0.438	0.66
ACgCL	4104 ± 787	4075 ± 689	0.312	0.76
EntR	2189 ± 365	1983 ± 373	4.412	< 0.001
EntL	2050 ± 399	1844 ± 356	4.240	< 0.001
MCgCR	4176 ± 547	4200 ± 541	–0.341	0.73
MCgCL	3988 ± 493	4002 ± 559	–0.213	0.83
MFCR	1581 ± 342	1505 ± 524	1.805	0.07
MFCL	1566 ± 285	1548 ± 291	0.487	0.62
OpIFGR	2575 ± 608	2424 ± 546	2.021	0.04
OpIFGL	2465 ± 550	2361 ± 579	1.466	0.14
OrIFGR	1252 ± 315	1196 ± 362	1.322	0.18
OrIFGL	1514 ± 335	1398 ± 356	2.658	< 0.001
PCgCR	3679 ± 466	3528 ± 415	2.657	< 0.001
PCgCL	3991 ± 442	3789 ± 424	3.676	< 0.001
PCuR	10129 ± 1193	9862 ± 1313	1.701	0.09
PCuL	10005 ± 1263	9759 ± 1299	1.522	0.13
SPLR	8867 ± 1140	8693 ± 1219	1.180	0.02
SPLL	8880 ± 1192	8662 ± 1313	1.390	0.17

Values are shown as mean ± standard deviation or percentage. Test statistics and *p*-values for differences between MCI-C and MCI-S are based on *t*-test. MCI-S, non-progressive MCI; MCI-C, progressive MCI; HippoR, Right Hippocampus; HippoL, Left Hippocampus; flWMR, frontal lobe WM right; flWML, frontal lobe WM left; plWMR, parietal lobe WM right; plWML, parietal lobe WM left; tlWMR, temporal lobe WM right; tlWML, temporal lobe WM left; ACgCR, Right ACgG anterior cingulate gyrus; ACgCL, Left ACgG anterior cingulate gyrus; EntR, Right Ent entorhinal area; EntL, Left Ent entorhinal area; MCgCR, Right MCgG middle cingulate gyrus; MCgCL, Left MCgG middle cingulate gyrus; MFCR, Right MFC medial frontal cortex; MFCL, Left MFC medial frontal cortex; OpIFGR, Right OpIFG opercular part of the inferior frontal gyrus; OpIFGL, Left OpIFG opercular part of the inferior frontal gyrus; OrIFGR, Right OrIFG orbital part of the inferior frontal gyrus; OrIFGL, Left OrIFG orbital part of the inferior frontal gyrus; PCgCR, Right PCgG posterior cingulate gyrus; PCgCL, Left PCgG, posterior cingulate gyrus; PCuR, Right PCu precuneus; PCuL, Left PCu precuneus; SPLR, Right SPL superior parietal lobule; SPLL, Left SPL superior parietal lobule.

Logit, the natural logarithm of the odds, is the key concept that underlies logistic regression. The equation for LR is:

$$\log \frac{P(y_i = 1 \,|x_i; \beta)}{1 - P(y_i = 1 \,|x_i; \beta)} = \sum_{j=1}^{N} \beta_j x_{ij} \qquad (2)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots \beta_N)^T$ are the parameters or weights of the logistic regression model, $x_{i,j} = (x_{i1}, \ldots x_{iN})$, $i = 1, \ldots M$. Also, $P(y_i = 1 \,|x_i; \boldsymbol{\beta})$ is the probability that *ith* MCI patient will develop dementia and $P(y_i = 0 \,|x_i; \boldsymbol{\beta})$ is the probability that *ith* MCI patient will not develop dementia. Denote $P(y_i = 1 \,|x_i; \boldsymbol{\beta}) = h(x_i)$, then

$$h(x_i) = \frac{1}{1 + \exp\left(\sum_{j=1}^{N} -\boldsymbol{\beta}_j x_{ij}\right)} \qquad (3)$$

LR is usually trained by minimizing an error function; an appropriate choice of such a function for binary classification problems is the cross-entropy error:

$$e_i(\boldsymbol{\beta}) = -y_i \log(h(x_i)) - (1 - y_i) \log(1 - h(x_i)) \qquad (4)$$

The total cost over the data $D = (x^i, y^i), i = 1,...M$ is:

$$J(\boldsymbol{\beta}) = -\frac{1}{M} \left[\sum_{i=1}^{M} y_i \log(h(x_i)) - (1 - y_i) \log(1 - h(x_i))\right] \quad (5)$$

Consider the problem of finding the maximum likelihood estimate (MLE) of the parameters β for the unregularized logistic regression model. To find the optimized weights β, the total cost needs to be minimized. The optimization function can be written:

$$\boldsymbol{\beta}^{optimal} = min_{\beta} - \frac{1}{M} \left[\sum_{i=1}^{M} y_i log(h(x_i)) - (1 - y_i) log(1 - h(x_i))\right] \quad (6)$$

Solving Equation (6) yields the optimal weights of $\boldsymbol{\beta}$. However, the model-building challenge is to abstract the underlying distribution from the particular instance D of samples because of the relatively small sample size, as compared to the number of features. The problem of replicating the data set instead of identifying the underlying distribution is known as overfitting [27]. To avoid the overfitting problem, it is often necessary to apply a dimension reduction technique. L_1 and L_2 norm are widely used to avoid overfitting, especially when there is a only small number of training examples, or when there is a larger number of features to be learned. L_1 norm or *lasso* is also often used for feature selection and has been shown to generalize well in the presence of many irrelevant features [28, 29]. L_1 regularization is implemented by adding L_1 norm to the cost function; the cost function and the optimization function were based on the following:

$$J(\boldsymbol{\beta}) = -\frac{1}{M} \left[\sum_{i=1}^{M} y_i log(h(x_i)) - (1 - y_i) log(1 - h(x_i))\right] + \lambda |\boldsymbol{\beta}| \quad (7)$$

and

$$\boldsymbol{\beta}^{optimal} = min_{\beta} \left\{-\frac{1}{M} \left[\sum_{i=1}^{M} y_i log(h(x_i)) - (1 - y_i) log(1 - h(x_i))\right] + \lambda |\boldsymbol{\beta}|\right\} \quad (8)$$

where λ is positive tuning parameter. This Equation (8) is referred to as L_1 regularized logistic regression.

Support vector machine

SVM is another classification and regression method that can handle high dimensional feature vectors. Algorithmically, SVMs build optimal boundaries between data sets by solving a constrained quadratic optimization problem [30–34]. The number of studies applying SVM to evaluate classification of conversion from MCI to dementia has grown over the past decade [1, 2, 5, 8, 23, 35–39].

We briefly review basic support vector machines with linear kernel (SVM-linear) for classification problems: Let $\boldsymbol{\beta}^T h(x) + \boldsymbol{\beta}_0 = 0$ denote an equidistant hyperplane (decision surface) to the closest point of each class on the new space. The goal of SVMs is to find $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ such that $\left|\boldsymbol{\beta}^T h(x) + \boldsymbol{\beta}_0\right| = 1$ for all points closer to the hyperplane. In the following classifier construction, one assumes that:

$$\boldsymbol{\beta}^T h(x_i) + \boldsymbol{\beta}_0 = \begin{cases} \geqslant 1 \; if \; y_i = 1 \\ \leqslant -1 \; if \; y_i = 0 \end{cases} \quad (9)$$

such that the distance from the closest point of each class to the hyperplane is $1/\|\boldsymbol{\beta}\|$ and the distance between the two groups is $2/\|\boldsymbol{\beta}\|$. To maximize the margin, the SVM requires the solution of the following optimization primal problem [40]:

$$min_{\beta,\beta_0} \sum_{i=1}^{M} \left\{1 - y_i \left[\beta_0 + \sum_{j=1}^{N} \beta_j^T h_j(x_{ij})\right]\right\} \quad (10)$$

where h_j is the kernel function which is a linear function for SVM-linear. Specifically, we choose, $h_j(x_j) = x_j$ for j-th covariate.

To make the algorithm work for highly correlated features and improve the fitted model's prediction accuracy, we reformulate our optimization by adding L_1-norm of β, i.e., the *lasso* penalty as follows:

$$min_{\boldsymbol{\beta},\boldsymbol{\beta}_0} \sum_{i=1}^{M} \left\{1 - y_i \left[\beta_0 + \sum_{j=1}^{N} \beta_j^T h_j(x_{ij})\right]\right\} + \lambda \|\beta\|_1 \quad (11)$$

where λ is the tuning parameter that controls the trade-off between loss and penalty. The lasso penalty shrinks the fitted coefficients $\boldsymbol{\beta}$ towards zero, and hence benefits from the reduction in fitted coefficients' variance.

Experimental design

We built four different classifiers, each designed to classify individual ADNI participants as belonging to

either the MCI-C group or the MCI-S group: Classifier 1 is logistic regression (C-LR); Classifier 2 is logistic regression with L_1 norm (C-LR-1); Classifier 3 is support vector machine (C-SVM); and Classifier 4 is SVM with L_1 norm (C-SVM-1). To test the classifiers' performance, we constructed five different data sources (Table 4). The first three single-modality data sets included clinical cognitive assessment scores and *APOE4* status (CCA), all MRI volumes (ROI-NP), and MRI volumes with preselection (ROI-P), respectively. Two additional multi-modal data sets were constructed by combining the CCA data separately with ROI-NP and ROI-P data sets (i.e., brain volumes with and without preselection). Furthermore, it is interesting to note that the number of MCI-S subjects is 101 (38%) in the Group One and 122 (39%) in Group Two, which makes the data rather imbalanced. Consequently, to precisely report the results obtained from the models, the present study also assessed additional model performance parameters, including AUC score, sensitivity, and specificity (accuracy coefficient is unreliable for imbalanced data). The prediction procedure consisted of three processing stages for Group One (Time = 36 months) and Group Two (Time = 24 months): 1) Split data as training, validation, testing set; 2) Train classifiers using training set, tune hyper-parameter using the validation set, and assess classifiers using testing set, then train classifiers again using L_1 norm on the same training set; 3) Report the testing accuracy, AUC score, sensitivity and specificity of each classifier on single-modality data. Specifically, the first stage used 80% of the sample as a training set while the remaining 20% of the data constituted the testing set. In the second stage, the optimal subsets of features of each data source are determined and chosen following application of L_1 norm. We then list the top 10 features of each data set for each of the models. In the last stage, we report AUC score, sensitivity (percent of MCI-C subjects correctly classified), and specificity (percent of MCI-S subjects correctly classified) as measures of classification accuracy. To protect against over-fitting and to avoid optimistically-biased estimates of model performance, we report 20 measures of predictive performance for each classifier (1–4); for these different partitions of the data, we report the mean and standard deviation of testing accuracy, AUC score, sensitivity, and specificity (Tables 6 and 7). We also investigate the relationship between the number of features and model performance. Finally, we compare the performance of LR with SVM based on their ability to handle the problem with a large number of covariates. Figure 3 illustrates the diagram of the prediction framework.

Table 4
Modalities

Data sources	# features
Single-modality	
Clinical Cognitive Assessments score and *APOE4* data (CCA)	19
ROI with no pre-selection data (ROI-NP)	259
ROI with pre-selection data (ROI-P)	26
Multi-modal	
CCA and ROI with no pre-selection data (CCAR-NP)	278
CCA and ROI with pre-selection data (CCAR-P)	45

RESULTS

Cross-validation and choice of λ

We adopted 10-fold cross-validation to tune the hyper-parameters for each model, which included dividing the data into separate sets for training and validation. The ratio of case in training and validation was 8:2. Here, the training set was used to train the model and the validation set was used to select the hyper-parameters. The results of a 10-fold cross-validation run are summarized with the mean and standard deviation of the model skill scores based on testing data. Cross-validation was also applied to tune the hyper-parameters; λ is used to denote the hyper-parameters for both LR-L_1 and SVML_1. To select the optimized λ, we tried different values of the λ; results reported here include values of λ = 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8 and applied them to the Eq (8) and (11). Next, we selected the λ value based on the best cross-validation score and used the selected λ with Classifiers 2 and 4 to select optimal features. For brevity, the model performance estimates are reported in Tables 6 and 7 for each different modalities, and the top 10 selected features are reported in Table 5. For example, the best λ for ROI-NP-L_1 was 0.01 and the top 3 optimal features selected by LR were left amygdala, right accumbens area, and right middle temporal gyrus. After hyper-parameters were selected, we adopted a 10-fold cross-validation again to avoid optimistically-biased estimates of model performance. In each iteration, 212 of the 265 participants are selected by simple random sampling as training cases and the remaining 53 were used as test cases. The approximate 4:1 ratio of training to test cases is, of course, arbitrary.

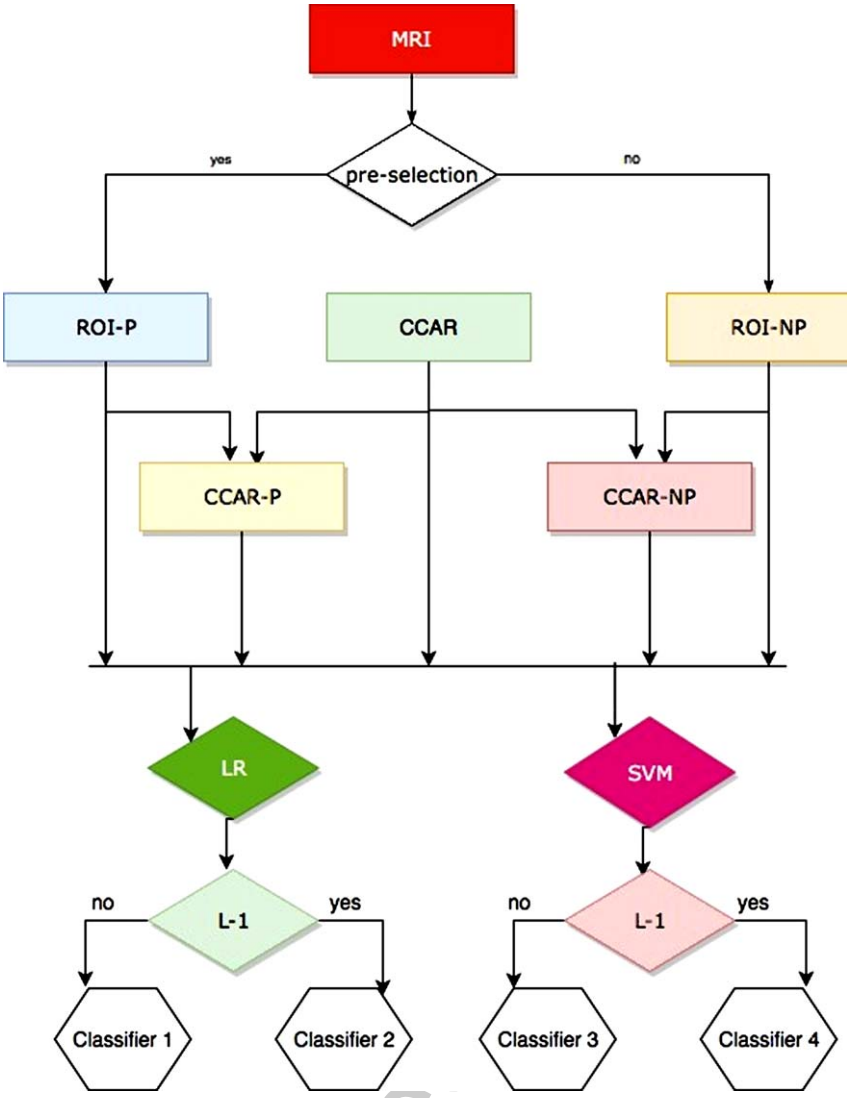


Fig. 3. Flowchart of the LR and SVM method. A) ROI-P: ROI level data with Pre-selection; B) ROI-NP: ROI level data with No Pre-selection; C) CCAR: Clinical, Cognitive assessments score, *APOE4*, and ROI level data.

Comparison with different modalities

We compared the performance of each classifier (1–4) on the five different feature sets (Table 4) based on estimates of AUC, sensitivity, and specificity. As shown in Table 6, the results of using LR with L_1 regularization (Classifier 2) can achieve the high AUC of 81.2% and sensitivity of 81.4% on single-modality data (CCA), which is considerably better than performance of LR on the other four modalities. Similarly, the best AUC and sensitivity achieved by SVM are 81.4% and 81.6% based on the combination of CCA and SVM-L1. Furthermore, we also found the highest accuracy achieved by both classifiers without applying regularization is based on the single-modality data (CCA); this indicated both classifiers perform best on single-modality data.

Comparison of pre-selection and L_1 norm

We found that using prior knowledge to inform feature selection improves model performance and protects against over-fitting. As shown in Table 6, model performance (i.e., AUC) on ROI-P (64.3%) and CCAR-P (76.3%) outperformed ROI-NP (60.6%) and CCAR-NP (60.1%). However, the performance of Classifier 2 on the ROI-NP-L_1 and CCAR-NP-L_1 data sets had AUC score of 64.1%

Table 5
Top 10 features of Group One obtained by L_1 regularization

Source	LR-L1 (Classifier 2)			SVM-L1 (Classifier 4)		
Data	CCA	ROI-NP	CCAR-NP	CCA	ROI-NP	CCAR-NP
1	FAQ	AmyL	FAQ	FAQ	AmyL	FAQ
2	mPACCtrailsB	AccmR	AmyL	Yrs. Educ.	AccmR	AmyL
3	*APOE4*	MTGR	ADASQ4	APOE4	AOrGL	AccmR
4	ADASQ4	HippoL	HippoL	mPACCdigit	PCgGL	AOrGL
5	Learning	AOrGL	MTGR	ADASQ4	HippoL	PTR
6	Yrs. Educ.	PrGR	APOE4	Learning	PrGR	AnGR
7	Forgetting	PCgGL	AOrGL	ADAS11	POrGR	APOE4
8	mPACCdigit	InfR	Learning	mPACCtrailsB	PTR	PCgGL
9	ADAS13	POR	mPACCtrailsB	DELTOTAL	LOrGL	Learning
10	ADAS11	MOGL	mPACCdigit	Forgetting	MOrGL	POrGR

AccmR, Right Accumbens Area; AmyL, Left Amygdala; HippoL, Left Hippocampus; InfR, Right Inf Lat. Vent; AOrGL, Left anterior orbital gyrus; AnGR, Left angular gyrus; LOrGL, Left lateral orbital gyrus; MOGL, Left middle occipital gyrus; MOrGL, Left medial orbital gyrus; MTGR, Right middle temporal gyrus; PCgGL, Left posterior cingulate gyrus; POR, Right parietal operculum; POrGR, Right posterior orbital gyrus; PrGR, Right precentral gyrus; PTR, Right planum temporal.

Table 6
LR and SVM performance of Group One (Time = 3 years) for models on single and multi-modal feature sets

Source	LR (Classifier 1 and 2)				SVM (Classifier 3 and 4)			
Modality	Test Acc%	AUC%	Sp%	Sn%	Test Acc%	AUC%	Sp%	Sn% #Features
CCA	74.3 ± 6.0,	80.8 ± 7.0,	62.3 ± 12.1	81.5 ± 6.2	72.4 ± 6.9,	80.0 ± 7.3,	53.6 ± 13.2	79.4 ± 7.7, $19^{(1)};19^{(2)}$
ROI-NP	58.1 ± 7.0,	60.6 ± 8.1,	45.5 ± 13.4	65.3 ± 7.9	59.5 ± 7.3,	61.4 ± 8.5,	46.5 ± 11.5	67.3 ± 8.5, $259^{(1)};259^{(2)}$
ROI-P	64.4 ± 6.5,	64.3 ± 6.6,	46.1 ± 10.4	75.0 ± 9.6	62.1 ± 5.9,	64.1 ± 6.2,	43.6 ± 9.5	78.4 ± 10.4, $26^{(1)};26^{(2)}$
CCAR-NP	57.6 ± 7.2,	60.1 ± 8.1,	44.8 ± 12.5	65.1 ± 9.0	57.8 ± 6.8,	59.1 ± 7.0,	45.9 ± 10.4	65.1 ± 7.5, $278^{(1)};278^{(2)}$
CCAR-P	72.7 ± 6.4,	76.3 ± 6.5,	60.5 ± 10.4	80.4 ± 8.2	66.9 ± 6.0,	69.2 ± 6.4,	53.6 ± 13.2	74.4 ± 10.5, $45^{(1)};45^{(2)}$
CCA-L_1	74.9 ± 6.4,	81.2 ± 6.7,	61.3 ± 12.0	83.1 ± 6.6	72.4 ± 6.0,	81.4 ± 6.9,	61.6 ± 11.5	81.6 ± 5.9, $4^{(1)};3^{(2)}$
ROI-NP- L_1	62.2 ± 6.6,	64.1 ± 7.9,	53.1 ± 13.1	68.1 ± 7.2	62.7 ± 5.8,	67.0 ± 6.7,	53.7 ± 11.6	67.7 ± 7.4, $29^{(1)};27^{(2)}$
ROI-P- L_1	64.4 ± 6.5,	64.3 ± 6.2,	46.2 ± 11.0	74.9 ± 9.6	64.4 ± 5.7,	64.7 ± 5.8,	46.7 ± 11.1	75.4 ± 8.3, $5^{(1)};17^{(2)}$
CCAR-NP- L_1	62.6 ± 7.2,	64.0 ± 8.2,	51.8 ± 12.7	69.5 ± 7.3	67.4 ± 6.4,	74.0 ± 7.4,	55.7 ± 12.1	74.1 ± 7.1, $18^{(1)};27^{(2)}$
CCAR-P- L_1	73.1 ± 6.5,	77.9 ± 5.9,	61.6 ± 10.5	79.6 ± 7.7	73.5 ± 6.2,	78.5 ± 6.4,	61.6 ± 9.3	80.8 ± 7.5, $14^{(1)};25^{(2)}$

Predictive performance of LR and SVM (mean ± standard deviation) for all models. Performance estimates include testing accuracy (Test Acc %), area under the cureve (AUC), sensitivity (Sn), and specificity (Sp). The number (#) of features was determined via (1): Classifier 2; (2): Classifier 4.

and 64.0%, while the ROIP-L_1 and CCAR-P-L_1 had respective AUC scores of 64.3% and 77.9%; this suggests that user-guided pre-selection significantly improved model performance over L_1 norm. In addition, the SVM (Classifiers 3 & 4) had similar and comparable results with LR classifiers. First, as with the LR models, the observed AUC estimates for CCAR-P and ROI-P (69.2% and 64.1%, respectively), were superior to AUCs from the CCAR-NP (59.1%) and ROI-NP analyses (61.4%). Classifier 4 exhibited similar performance on the CCAR-P-L_1 as Classifier 2, with an AUC value of 79.6%—higher than the model for CCAR-NPL_1 (74.0%). Therefore, manually selecting features improves model's performance whether L_1 norm is applied, or not. Second, these results show it is necessary and important to use pre-selection because both LR and SVM models on CCAR-P-L_1, with respective AUC estimates of 77.9% and 78.5%, exhibited superior performance over the models without such pre-selection (i.e., LR and SVM on CCAR-NP-L_1 had AUC estimates of 64.0% and 74.0%, respectively).

Comparison of groups one and two

In addition to the results from models of Group One (i.e., MCI-to-AD conversion over 36 months), we also evaluated the performance of Group Two (i.e., MCI-to-AD conversion over 24 months) in an effort to gain further insight regarding possible benefits of shorter or longer assessment periods on classification of the progression of MCI to dementia. Table 7 summarizes the predictive performance of LR and SVM for Group Two. Similarly, we also evaluated classifier performance for single- and multi-modality feature sets. The best result is obtained by using SVM-L_1 model (Classifier 4) on CCAR-P, and its corresponding AUC, Sn and Sp are 76.2%, 60.1%, and

Table 7
LR and SVM performance of Group Two (Time = 2 years) for single-data and multi-modal data

Source	LR (Classifier 1 and 2)				SVM (Classifier 3 and 4)			
Modality	Test Acc%	AUC%	Sp%	Sn%	Test Acc%	AUC%	Sp%	Sn% #Features
CCA	69.9 ± 5.3,	76.2 ± 5.5,	56.7 ± 9.0	79.3 ± 7.3	69.4 ± 5.4,	75.4 ± 5.5,	56.7 ± 8.8	78.6 ± 7.1, $19^{(1)}$; $19^{(2)}$
ROI-NP	58.1 ± 4.2,	58.8 ± 5.6,	49.7 ± 7.1	64.4 ± 5.9	57.8 ± 5.0,	56.6 ± 6.4,	50.3 ± 7.1	62.9 ± 7.5, $259^{(1)}$; $259^{(2)}$
ROI-P	63.4 ± 4.7,	65.8 ± 4.3,	43.7 ± 10.2	77.8 ± 8.6	64.5 ± 4.7,	66.2 ± 5.0,	44.5 ± 8.5	79.1 ± 9.1, $26^{(1)}$; $26^{(2)}$
CCAR-NP	57.3 ± 4.0,	58.8 ± 5.4,	47.5 ± 8.3	64.3 ± 5.8	56.6 ± 5.5,	56.4 ± 5.2,	48.9 ± 7.9	62.3 ± 10.4, $278^{(1)}$; $278^{(2)}$
CCAR-P	70.2 ± 5.4,	74.0 ± 5.0,	56.7 ± 9.5	80.6 ± 7.0	69.5 ± 4.9,	72.0 ± 5.3,	58.1 ± 8.1	78.0 ± 8.2, $45^{(1)}$; $45^{(2)}$
CCA-L_1	70.1 ± 4.8,	76.3 ± 5.3,	56.8 ± 9.9	79.8 ± 7.6	70.4 ± 4.9,	76.4 ± 7.7,	56.8 ± 9.8	79.4 ± 7.7, $4^{(1)}$; $3^{(2)}$
ROI-NP- L_1	62.2 ± 6.0,	64.7 ± 6.0,	48.8 ± 9.2	72.0 ± 6.8	60.8 ± 4.5,	65.9 ± 6.1,	53.6 ± 7.5	64.3 ± 7.9, $29^{(1)}$; $31^{(2)}$
ROI-P- L_1	64.1 ± 4.6,	66.8 ± 3.8,	42.8 ± 11.3	79.8 ± 8.4	65.4 ± 4.0,	67.8 ± 3.9,	46.3 ± 9.4	81.8 ± 7.2, $6^{(1)}$; $14^{(2)}$
CCAR-NP- L_1	62.6 ± 6.3,	64.8 ± 6.0,	49.1 ± 9.1	72.1 ± 6.1	64.5 ± 5.1,	71.7 ± 4.8,	55.4 ± 7.8	71.4 ± 8.9, $26^{(1)}$; $32^{(2)}$
CCAR-P- L_1	70.0 ± 5.5,	74.3 ± 5.5,	57.8 ± 8.0	78.3 ± 8.8	71.3 ± 4.9,	76.2 ± 4.7,	60.1 ± 7.1	79.2 ± 8.5, $14^{(1)}$; $27^{(2)}$

For each modality, the predictive performance of LR and SVM are shown (mean $\pm$ standard deviation), including testing accuracy, AUC, sensitivity (Sn), specificity (Sp), # features is the number of features; # features is the number of features; this parameter was determined via (1): Classifier 2; (2): Classifier 4.

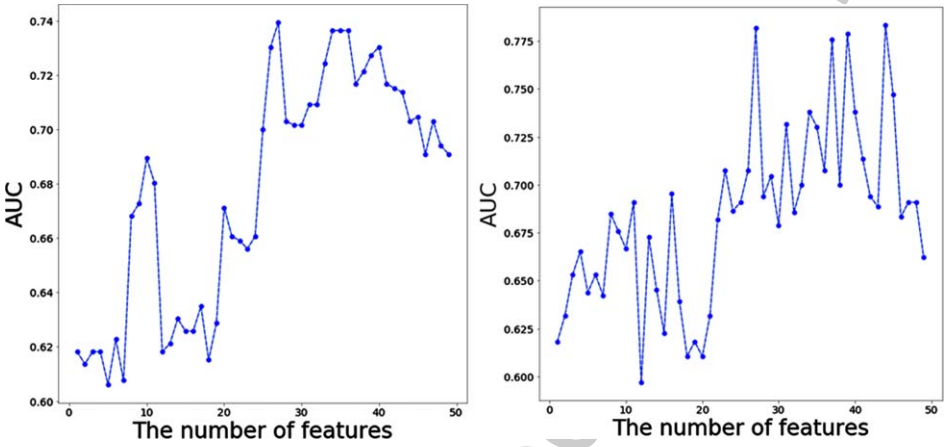


Fig. 4. Model performance on ROI feature set by number of features for LR and SVM. Panel (a) shows dramatic growth in AUC with LR as the number of features increases from 1 to 30, and then becoming more static at approximately 74%, i.e., as the number of features increases from 30 to 40, but drops significantly when the number of features reaches to 41. Panel (b) shows the AUC increased dramatically as the number of features grows from 1 to 28, but fluctuated after 29. The optimal number of ROI features for both methods are 29 and 28, and their corresponding optimized AUC were approximately 74.0% and 78.0%.

79.2%, which verifies the assumption that manually selecting techniques improves the model's performance again. However, it warrants mention that all classifiers' performance on the Group One data outperformed the same classifiers' performance on the same data sets in Group Two. For example, Classifier 2 of Group One on CCA achieved AUC and Sn values of 81.2% and 83.1%, which is considerably better than the same classifier of Group Two on CCA (i.e., 76.3% and 79.8%). Similarly, Classifier 3 for ROI-NP had an AUC of 61.4% for Group One and 56.6% for Group Two. The experimental results indicated superior model performance on data obtained using longer than using shorter follow-up periods. Given the uncertainty in conversion, a longer time window for assessment of cognitive and functional change clearly yields more accurate classification.

Comparison of LR and SVM

In addition to comparing classification between different time windows of assessment, we also compared performance differences between LR and SVM. The results, including models' ability to address the overfitting problem of LR and SVM methods with different modalities are displayed in Tables 6 and 7 and Figs. 4 and 5. First, it is worth noting that both LR and SVM do not work well if no L_1 penalization used, since Classifiers 2 and 4 outperform Classifiers 1 and 3 on the same data set. Second, it

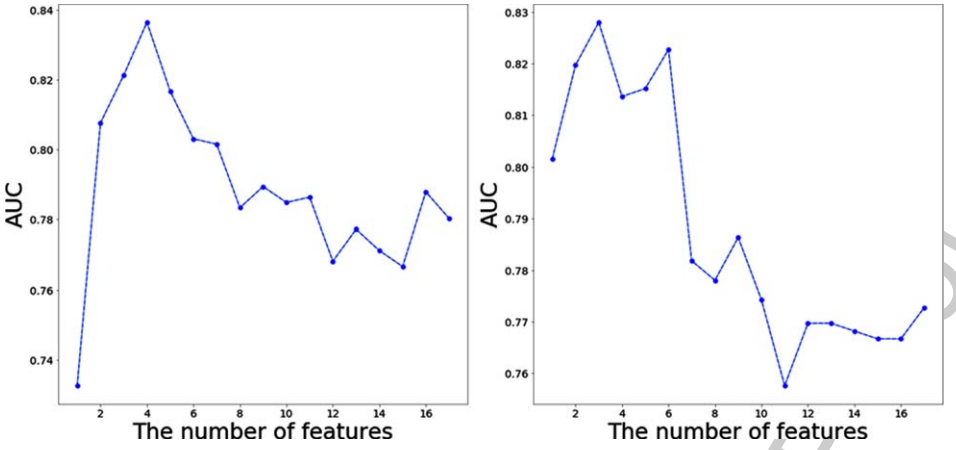


Fig. 5. Model performance on CCA feature set by number of features for LR and SVM. Figure (a) shows there is a significant increase in the AUC with LR as the number of features increases from 1 to 5, then there is a slight decrease in the testing accuracy when the number of features is greater than 5. Figure (b) shows the AUC shot up dramatically as the number of features increases from 1 to 4. The optimal number of CCA features obtained by LR and SVM are 5 and 4, and their corresponding optimized AUC are approximately 84.0% and 83.0%.

is worth noting that SVM has a better performance on MRI data when the L1 feature selection method is employed. Third, it was possible to obtain good performance accuracy using LR, which had equivalent model performance as SVM for "large p" data (ROI-P), as evidenced by respective AUC estimates for Classifiers 1 and 3 of 64.3% and 64.1%. Finally, as shown in Figs. 5 and 4, the SVM method is more stable and robust than LR to the large number of features when n is small. To summarize, the best performance of Group One was achieved by Classifier 4 (SVM with L_1 norm) when using multi-modal, i.e., CCAR-L1, had an AUC of 81.4%.

DISCUSSION

In this study, we applied two machine learning methods under multiple conditions, to test accuracy in classifying patients with MCI who progress to clinically-defined dementia (MCI-C) from those who remain stable (MCI-S). Using multi-modal data from ADNI, we compared LR and SVM classification accuracy and pre-selection dimensional reduction techniques, i.e., feature selection as informed by prior findings in clinical neuroscience and by L_1 norm. Notably, the present results demonstrate important boundaries for applying feature selection techniques in statistical classification of MCI-to-dementia conversion. Specifically, we found that while using L_1 for pre-selection can improve accuracy, it also benefits from a more limited, theoretically based set of feature inputs. In addition, we found that model performance benefited from a longer window of assessment. These results have implications for studies utilizing multi-modal data for such classification, including features from clinical neuropsychological assessment, demographic and genetic markers, MRI-based volumetric brain measures, and other modalities.

Comparison of user-defined and L_1 pre-selection for LR and SVM classifiers yielded multiple noteworthy findings, consistent with previously published reports [1, 2, 5, 8, 13, 23, 25, 35]. First, the classification results showed that the model using multi-modal data with cognitive, clinical, and volumetric data (CCAR) achieved better classification accuracy than the methods based on single-modality (CCA, ROI). Moreover, the AUC of CCAR based on LR or SVM was either statistically significantly or at least numerically greater than those based on the single-modality model. Based in AUC, we reported the highest accuracy was observed for CCAR data at 78.5% by L_1 SVM and 77.9% by L_1 LR. Second, SVM demonstrated several advantages over LR in discriminating MCI-C from MCI-S (Fig. 4). For one, SVM performance tended to be more stable than LR when the number of features was relatively large. In other words, the model performance of SVM on ROI data remained more stable than LR when using larger numbers of features without user-defined pre-selection. In particular, SVM performance on ROI data improved as the number of features increased from 20 and 30. In contrast, the AUC values for ROI data sets remained fairly static despite increasing the number of features. However, LR model performance decreased gradually after the number of ROI features reached 40. Third, the classification results clearly

demonstrate that manually selecting features on MRI data not only improved the model performance and protected the classifier from overfitting, but also affords easier interpretation of each selected feature's contribution to the model. In addition, we show that pre-selection improves performance: Tables 6 and 7 suggest it is the best strategy to obtain the maximum model performance, compared to features selection based on L_1 norm.

The present findings can also be interpreted in the context of other reports over the past decade that also investigated the prognostic capacity of brain volumetry data to predict the conversion of MCI to dementia, using either SVM or LR, and that also combined volumetry data with other imaging and biomarker modalities such as MRI, functional MRI (fMRI), PET to cerebrospinal fluid (CSF) protein markers [1, 2, 5, 8, 13, 23, 25, 35, 41–43]. In addition, one can vary the degrees of non-linearity and flexibility in the model by employing different kernel functions. For example, Young et al (2013) report [8], results from both SVM and Gaussian process (GP) classification on MCI progression in ADNI data using MRI, PET, *APOE4*, and CSF biomarkers. In contrast the present study and with other published work that used MCI-C and MCI-S groups as training and test data sets, they trained a classifier to distinguish cognitively normal older adults from those diagnosed as probable AD. They reported that the accuracy using GP, an AUC value of 79.5%, was substantially higher than using any individual modality or using multi-kernel SVM. Other studies of MCI-to-dementia classification reporting high accuracy have also implemented other approaches such as multiple kernel learning (pMKL) classification techniques using clinical, MRI and plasma biomarkers data. One method using this approach to identify the important features first grouped the data set into five different data sources and then applied a filter-wrapper approach of feature selection techniques in combination with Joint Mutual Information (JMI) criterion to achieve an AUC of 82% [23].

We also found consistently superior classification performance in patients classified under a longer window of assessment. MCI-to-dementia conversion is a process that can take several years to reliably track an individual from onset of amnestic MCI to early-stage dementia [8, 44, 45]. For the modeled features to be of use for classification necessitates well-defined, if not orthogonal classes. However, MCI is not inherently prodromal to dementia: a large proportion of individuals with MCI never progress, either reverting to cognitively normal status or remaining rather stable. Furthermore, others may show early evidence of brain atrophy that precedes cognitive impairment by years. In order to account for this variable timing, others have employed methods such as supervised learning using time windows [46]; however, even those methods strongly benefit from longer follow-up periods. Thus, MCI is an inherently heterogeneous and poorly-defined class, particularly in terms of the relationships between brain characteristics and the likelihood and timing of further cognitive decline.

The brain volumetric data evaluated in the present study were to limited baseline MRI scans. Alternatively, classifying cognitive decline may benefit from further extending the model to accommodate repeated measurements from longitudinal data. While the inclusion of repeated volumetric data should improve classification accuracy, quantifying the improvements in model performance may also depend on other factors, such as added noise or redundancy from additional brain parameters, or variability in disease progression. In addition, most recent computational neuroimaging studies in the past few years have utilized features from multiple neuroimaging modalities [5, 26, 36, 37, 39, 47–50]. For example, when Ding et al. applied SVM with PET and MRI data to classify the transition from MCI to AD, they reported the sensitivity and specificity were 66.67% and 64.52% [36]. In addition to PET and structural MRI data, CSF protein markers can be used to predict progression from MCI to AD, in addition to proteomic, demographic, and cognitive data [38, 51, 52]. By applying LR with L_1 norm to CSF markers for classifying individual patients as belonging to either the MCI-C and MCI-S group, one study reported a sensitivity and specificity of 80% and 75% [26]. Furthermore, Varatharajah and colleagues (2020) showed SVM-linear outperforms other advanced classification methods, including linear classifiers—multiple kernel learning (MKL) with linear kernels, SVM with a linear kernel, and generalized linear model (GLM), in predicting transition from MCI to AD [42]. In general, LR works well when the data is linearly separable and the number of data is greater than the number of features, whereas SVM with Gaussian Kernel is mostly used when the data is not linearly separable. In addition to LR and SVM, deep neural network approaches also offer benefits [41, 53], but have not had the extent of application in ADNI data as SVM and LR. Using a novel LR, artificial neural network (ANN) model and decision tree (DT) model for classifying the progression

of MCI to AD, Kuang (2021) reported that the ANN exhibited the highest sensitivity at 82.1% [43].

In conclusion, models applying prior knowledge for classification and prediction of MCI-to-dementia conversion outperform those without pre-selection. This theoretically guided pre-selection of features from MRI-based regional brain volumes appears to protect the model against over-fitting. In addition, the present findings demonstrate that SVM classifier performance is more stable than LR for dealing with the "large p" problem. Clinical researchers should both note the value of evaluating different classification and pre-selection approaches in application to clinical or research questions and be mindful that not all machine learning techniques are equally beneficial for modeling specific clinical outcomes.

ACKNOWLEDGMENTS

The research is partially supported by NSF-DMS 1945824 and 1924724.

We are grateful to the patients and their families who participated in the ADNI.

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. 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 is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://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.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-1398r3).

REFERENCES

[1] Zhang D, Shen D (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* **59**, 895-907.

[2] Zhang D, Shen D (2012) Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS One* **7**, 1-15.

[3] Korolev IO (2014) Alzheimer's disease: A clinical and basic science review. *Med Stud Res J* **4**, 24-33.

[4] Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ (2009) Mild cognitive impairment: Ten years later. *Arch Neurol* **66**, 1447-1455.

[5] Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T (2011) Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS One* **6**, e21896.

[6] Farlow MR (2009) Treatment of mild cognitive impairment (MCI). *Curr Alzheimer Res* **6**, 262-267.

[7] Salazar D, V´elez J, Salazar J (2012) A relationship between the transient structure in the monomeric state and the aggregation propensities of alpha-synuclein and beta-synuclein. *Biochemistry* **53**, 7170-7183.

[8] Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S (2013) Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *Neuroimage Clin* **2**, 735-745.

[9] Chen S, Hsiao Y, Huang Y (2009) Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power Doppler imaging. *Korean J Radiol* **10**, 464-471.

[10] Salazar D, Velez J, Salazar J (2012) Comparison between SVM and logistic regression: Which one is better to discriminate. *Rev Colomb Estad* **35**, 223-237.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825-2830.

[12] McKinney W (2010) Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, pp. 51-56.

[13] Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N (2012) Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. *BMC Neurol* **12**, 46.

[14] Chapman RM, Mapstone M, McCrary JW, Gardner MN, Porsteinsson A, Sandoval TC, Guillily MD, Degrush ED, Reilly LA (2011) Predicting conversion from mild cognitive impairment to Alzheimer's disease using

neuropsychological tests and multivariate methods. *J Clin Exp Neuropsychol* **32**, 187-189.

[15] Ewers M, Walsh C, Trojanowski JQ, Shaw LM, Petersen RC, Jack CR, Feldman HH, Bokde AL, Alexander GE, Scheltens P, Vellas B, Dubois B, Weiner M, Hampel H (2012) Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol Aging* **33**, 1203-1214.

[16] Tabatabaei-Jafari H, Shaw ME, Cherbuin N (2015) Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis. *Alzheimers Dement (Amst)* **1**, 487-504.

[17] 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.

[18] Eckerstr¨om C, Olsson E, Borga M, Ekholm S, Ribbelin S, Rolstad S, Starck G, Edman A, Wallin A, Malmgren H (2008) Small baseline volume of left hippocampus is associated with subsequent conversion of MCI into dementia: The Goteborg MCI study. *J Neurol Sci* **271**, 48-59.

[19] Risacher SL, Saykin AJ, West JD, Shen L, Firpi HA, McDonald BC (2009) Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr Alzheimer Res* **6**, 347-361.

[20] DoshivJ, Erus G, Rozycki M, Davatzikos C (2016) Hierarchical parcellation of MRI using multi-atlas labeling methods. *Alzheimer's Disease Neuroimaging Initiative*, Philadelphia, PA

[21] Doshi J, Erus G, Ou Y, Gaonkar B, Davatzikos C (2013) Multi-atlas skull-stripping. *Acad Radiol* **20**, 1566-1576.

[22] Doshi J, Erus G, Ou Y, Resnick S, Gur R, Satterthwaite T, Davatzikos C (2015) MUSE: Multiatlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage* **127**, 186-195.

[23] Korolev I, Symonds L, Bozoki A (2016) Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS One* **11**, 2.

[24] Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, Decruyenaere J (2009) Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med Inform Decis Mak* **8**, 56.

[25] Devanand DP, Liu X, Tabert MH, Pradhaban G, Cuasay K, Bell K (2008) Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biol Psychiatry* **64**, 871-879.

[26] Llano DA, Bundela S, Mudar RA, Devanarayan V (2017) A multivariate predictive modeling approach reveals a novel CSF peptide signature for both Alzheimer's Disease state classification and for predicting future disease progression. *PLoS One* **12**, 1-18.

[27] Stephan D, Lucila O (2008) Logistic regression and artificial neural network classification models: A methodology review. *BMC Med Inform Decis Mak* **8**, 56-64.

[28] Lee S, Lee H, Abbeel P, Ng Andrew (2006) Efficient L_1 regularized logistic regression. *Association for the Advancement of Artificial Intelligence*, pp. 401-408.

[29] Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Series B* **58**, 267-288.

[30] Cristianini N, Shawe-Taylor J (2000) *An introduction to support vector machines and other kernelbased learning methodsr*. Cambridge University Press, Cambridge, United Kingdom.

[31] Scholkopf B, Smola A (2002) *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, Boston.

[32] Vapnik V (1998) *Statistical learning theory*. John Wiley, New York.

[33] Vapnik V (1995) *The nature of statistical learning theory*. Springer-Verlag, New York.

[34] Vapnik V (1998) *The support vector method of function estimation*. Kluwer Academic Publisher, Boston, pp. 267-288.

[35] Hinrichs C, Singh V, Xu G, Johnson SC (2011) Predictive markers for AD in a multi-modal framework: An analysis of MCI progression in the ADNI population. *Neuroimage* **55**, 574-589.

[36] Ding J, Huang Q (2017) Prediction of MCI to AD conversion using Laplace Eigenmaps learned from FDG and MRI images of AD patients and healthy controls. *2nd International Conference on Image, Vision and Computing (ICIVC)*, pp. 660-664.

[37] Hojjati S, Ebrahimzadeh A, Khazaee A, Babajani-Feremi A (2017) Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *J Neurosci Methods* **282**, 69-80.

[38] Davatzikos C, Bhatt P, Shaw L, Batmanghelich K, Trojanowski J (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* **32**, 19-27.

[39] Wei R, Li C, Fogelson N, Li L (2016) Prediction of conversion from mild cognitive impairment to Alzheimer's disease using MRI and structural network features. *Front Aging Neurosci* **8**, 76.

[40] Zhu J, Rosset S, Hastie T, Tibshirani R (2003) *1-Norm Support Vector Machines*. MIT Press, pp. 49-56.

[41] Li F, Tran L, Thung K, Ji S, Shen D, Li J (2015) A robust deep model for improved classification of AD/MCI patients. *IEEE J Biomed Health Inform* **19**, 1610-1616.

[42] Varatharajah Y, Ramanan VK, Iyer R, Vemuri P (2019) Predicting short-term MCI-to-AD progression using imaging, CSF, genetic factors, cognitive resilience, and demographics. *Sci Rep* **2235**, 9.

[43] Kuang J, Zhang P, Cai T, Zou Z, Li L, Wang N, Wu L (2021) Prediction of transition from mild cognitive impairment to Alzheimer's disease based on a logistic regression-artificial neural network decision tree model. *Geriatr Gerontol Int* **21**, 43-47.

[44] Mitchell A, Shiri-Feshki M (2008) Temporal trends in the long term risk of progression of mild cognitive impairment: A pooled analysis. *J Neurol Neurosurg Psychiatry* **79**, 1386-91.

[45] Lee S, Bachman A, Yu D, Lim J, Ardekani B (2016) Predicting progression from mild cognitive impairment to Alzheimer's disease using longitudinal callosal atrophy. *Alzheimers Dement (Amst)* **2**, 68-74.

[46] Pereira T, Lemos L, Cardoso S, Silva D, Rodrigues A, Santana I, DeMendon¸ca A, Guerreiro M, Madeira SC (2017) Predicting progression of mild cognitive impairment to dementia using neuropsychological data: A supervised learning approach using time windows. *BMC Med Inform Decis Mak* **17**, 110.

[47] Shen T, Jiang J, Li Y, Wu P, Zuo C, Yan Z (2018) A multivariate predictive modeling approach reveals a novel CSF peptide signature for both Alzheimer's Disease state classification and for predicting future disease progression. *40th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 738-741.

[48] Menikdiwela M, Nguyen C, Shaw M (2018) Deep learning on brain cortical thickness data for disease classification. *Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-15.

[49] Minhas S, Khanum A, Riaz F, Alvi A, Khan SA (2017) A nonparametric approach for mild cognitive impairment to AD conversion prediction: Results on longitudinal data. *IEEE J Biomed Health Inform* **21**, 1403-1410.

[50] Wang B, Hong R, Xu Y, Zhou F, Wang P (2016) Identifying mild cognitive impairment conversion to Alzheimer's disease from medical image information. *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 1-2.

[51] Shaffer J, Petrella J, Sheldon F, Choudhury K, Calhoun V and Coleman R, Doraiswamy P (2012) Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. *Radiology* **266**, 583-591.

[52] Cheng B, Zhang D, Shen D (2012) Domain transfer learning for MCI conversion prediction. *Med Image Comput Comput Assist Interv* **15**, 82-90.

[53] Suk H. I, Shen D (2013) Deep learning-based feature representation for AD/MCI classification. *Med Image Comput* **6**, 583-590.