Article:

Bucholc, M, Titarenko, S orcid.org/0000-0002-4453-0180, Ding, X et al. (2 more authors) (2023) A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia. Expert Systems with Applications, 217. 119541. ISSN 0957-4174

https://doi.org/10.1016/j.eswa.2023.119541

A hybrid machine learning approach for prediction of conversion from mild cognitive impairment to dementia

Magda Bucholc[a*], Sofya Titarenko[b], Xuemei Ding[a], Callum Canavan[c], Tianhua Chen[d]

[a] Cognitive Analytics Research Lab, School of Computing, Engineering & Intelligent Systems, Ulster University, BT48 7JL, Northern Ireland, United Kingdom (E-mails: m.bucholc@ulster.ac.uk, x.ding@ulster.ac.uk)

[b] School of Mathematics, University of Leeds, LS2 9JT, United Kingdom (Email: s.titarenko@leeds.ac.uk)

[c] Seagate Technology, BT48 0LY, Northern Ireland, United Kingdom (E-mail: callum.canavan@seagate.com)

[d] University of Huddersfield, HD1 3DR, United Kingdom (E-mail: t.chen@hud.ac.uk)

*Corresponding author: Magda Bucholc; Cognitive Analytics Research Lab, School of Computing, Engineering & Intelligent Systems, Ulster University, BT48 7JL, Northern Ireland, United Kingdom. Phone: +44 28 7167 5155. E-mail: m.bucholc@ulster.ac.uk

Abstract

Mild cognitive impairment (MCI) represents a precursor to dementia for many individuals; however, some forms of MCI tend to remain stable over time and do not progress to dementia (Jicha et al., 2006; Petersen et al., 1999; Visser et al., 2006). In fact, conversion rates vary substantially depending on the diagnostic criteria used and the nature of the analytic sample and clinical setting (Ganguli et al., 2004; Ritchie, Artero, & Touchon, 2001). To identify personalized strategies to prevent or slow the progression of dementia and to support the clinical development of novel treatments, we need to develop new approaches for modelling disease progression that can differentiate between progressive and non-progressive MCI subjects. The aim of this study was to develop a novel prognostic machine learning (ML) framework utilising longitudinal information encoded in efficient, cost-effective, and non-invasive markers to identify MCI subjects that are at risk for developing dementia. Our approach was developed using the dataset from the National Alzheimer's Coordinating Center. We built two prognostic models based on the patient data from 3 (n=768) (Model 1) and 4 (n=409) (Model 2) assessment visits. A novel hybrid prognostic approach, using cognitive trajectory classes, generated through unsupervised learning (Stage 1), as input in supervised ML models (Stage 2), was developed and systematically tested. Our unsupervised learning approach (Stage 1) involved: (i) the implementation of the longitudinal data partitioning method allowing for clustering trajectories based on their shapes; (ii) validation of the optimal number of clusters using three different Clustering Validity Indices (CVIs), and (iii) application of the fusion-based methods for combining CVIs into the fused normalized CVI scores, averaged for each cluster partition to determine the final number of trajectory classes for each type of clinical scores. In Stage 2, we built four types of prognostic models based on random forest (RF), Support Vector Machines (SVM), logistic regression (LR), and kNN ensemble approaches. Classification models incorporating both clinical scores and cognitive trajectory classes input showed up to 6.5% higher accuracy than models based only on clinical scores ($p < 0.05$ in all cases). Given the patient data from three time points (Model 1), the highest recorded prediction accuracy was achieved for the ensemble and RF model, i.e., 85.0% (standard deviation: 3.1%) and 84.6% (4.1%) respectively. Using the patient data from four time points (Model 2), the highest accuracy was reported for RF and ensemble models, i.e., 87.5% (6.1%) and 86.8% (3.7%) respectively. We showed that the incorporation of the output of unsupervised learning significantly improved the performance of supervised ML models. Our prognostic framework can be applied to improve recruitment in clinical trials and to select early interventions for individuals at high risk of developing dementia.

Keywords: dementia; mild cognitive impairment; machine learning; longitudinal modelling; unsupervised learning, prognostic model

1 Introduction

The structural and function changes in the brain, associated with changes in neuronal structure, loss of synapses, and neuronal dysfunction, become more likely as people age . Neurodegenerative diseases, such as Alzheimer's disease and related dementias, accelerate cognitive decline, with many people developing cognitive impairment severe enough to affect their functional independence and social interaction (Murman, 2015). Worldwide, approximately 50 million people live with dementia, and this number is projected to triple by 2050 (Murman, 2015). Patients diagnosed with mild cognitive impairment (MCI) face a substantially higher risk of developing dementia (Prince et al., 2015). MCI is often considered a transitional phase between healthy ageing and dementia, with the annual rate of conversion of 9.6% (Mitchell & Shiri-Feshki, 2009). Nevertheless, not all MCI patients progress to dementia and some even regain normal cognition. Finding new approaches to differentiate between progressive and non-progressive MCI subjects and developing new methodologies for modelling disease progression is therefore of paramount importance.

Traditionally, the severity and changes in cognitive function have been assessed manually by clinicians using appropriate guidelines (e.g., NINCDS-ADRDA, DSM-IV, ICD-10). In recent years, advances in machine learning (ML) have provided the computational framework that has significant potential to revolutionize clinical decision making by leveraging abundant patient data and providing risk assessments and recommendations tailored to individual patients (Topol, 2019; Yu, Beam, & Kohane, 2018). As such, the data-driven identification of disease states has offered unique opportunities for enhancing disease classification based on expert knowledge (Yu et al., 2018). In the context of dementia, accurate prediction of future dementia-related cognitive decline has important practical applications. In particular, the identification of individuals at risk of developing dementia can help healthcare professionals make more informed clinical decisions on treatment strategies. It can also support the clinical development of novel treatments by creating the opportunity for those at increased risk of developing dementia to participate in preventive interventions and clinical trials and be assessed for potential contributors to cognitive impairment (e.g., vitamin deficiency, medication-side effects, psychiatric conditions, brain injuries).

The most common ML approaches used to assist the diagnosis of dementia have been based on supervised ML methods that learn the mapping function from the labelled data. This includes Random Forest (RF) (Bansal, Chhikara, Khanna, & Gupta, 2018; Bucholc et al., 2019; Gray et al., 2013; Sarica, Cerasa, & Quattrone, 2017), K-Nearest-Neighbors (KNN) (Bucholc et al., 2019; Varatharajan, Manogaran, Priyan, & Sundarasekar, 2018), Logistic Regression (LR) (Barnes et al., 2010; Bauer, Cabral, & Killiany, 2018), Naïve Bayes (Bansal et al., 2018; Shree & Sheshadri, 2018), Support Vector Machine (SVM) (Bucholc et al., 2019; Varatharajan et al., 2018), linear regression (Bauer et al., 2018), and fuzzy classifier systems (Stirling, Chen, & Bucholc, 2021). Most practical deep learning dementia applications have also been driven by supervised learning (Ding et al., 2019; Jo, Nho, & Saykin, 2019). In contrast, unsupervised learning has been rarely applied in dementia context, possibly because it involves more complex processing tasks. In unsupervised learning*,* no priori information exists, and the modelling process is based solely on identifying regularities in datasets (Celebi & Aydin, 2016). This lack of direction for the learning algorithm in unsupervised learning is in fact advantageous, since it overcomes limitations of the supervised feature space definition by unbiasedly revealing associations without any human involvement. Some examples of utilization of unsupervised ML methods in dementia research can be found in (de Langavant, Bayen, & Yaffe, 2018; Escudero, Zajicek, & Ifeachor, 2011). Although unsupervised machine learning approaches may be difficult to understand from a clinical perspective, their use does not depend on the availability of a prespecified clinical outcome such as, clinical diagnosis and hence, can be more easily re-applied using different types of datasets. In addition, when combined with supervised ML, unsupervised learning can exploit information of unlabelled data to improve the accuracy of supervised models. As such, the unsupervised learning can generate labels that can be used for a supervised learning task.

In this study, we developed a novel hybrid prognostic framework utilising longitudinal information encoded in cognitive scores to identify subjects with mild cognitive impairment (MCI) that are at risk for developing dementia. We first allowed for hypothesis-free detection of cognitive trajectory classes within the data without being guided by a pre-labelling of instances and then, combined this information with the routinely collected cognitive assessment scores to predict which MCI individuals are likely to develop dementia and which are suffering from non-progressive cognitive impairment. It was hypothesized that incorporation of cognitive trajectory classes, generated using unsupervised learning, as additional input variables would improve performance metrics of supervised ML models.

2 Material and methods

2.1 Participants

The data used in this study was taken from the National Alzheimer's Coordinating Center Uniform Data Set (NACC-UDS), containing participant characteristics collected in Alzheimer's Disease Research Centers (ADRC) in the period 2005-2018 (Beekly et al., 2004; Beekly et al., 2007). The primary purpose of the NACC-UDS is to provide a standard set of assessment procedures, collected longitudinally (approximately annually), to describe ADRC participants with AD and mild cognitive impairment and compare them to cognitively healthy controls (Morris et al. 2006). This longitudinal information is captured during the participants' initial visit and subsequent follow-up visits by trained research personnel, clinicians, and psychometrists. Participant recruitment occurs through referrals from neurologists and community outreach. The incidence of MCI and AD are determined based on the clinical diagnosis made by a single physician or a consensus panel according to each ADRC's diagnostic protocol; however, each ADRC generally adheres to the modified Petersen criteria for establishing MCI diagnoses and to the DSM-IV or NINDS-ADRDA guidelines for the clinical diagnosis of AD. The NACC-UDS includes demographic information, medical history, family history, medication usage, cognitive assessments, neurological exams, and clinical diagnoses (Morris et al., 2006). The authors assert that all procedures contributing to this work complied with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human subjects were approved by the University of Washington Institutional Review Board. Written informed consent was obtained from all subjects at each participating ADRC.

The NACC dataset contained 37,568 unique participants. Overall, the average number of visits per participant was 3.4 visits (min = 1, max = 13). Out of 37,568 participants, 12,231 (32.6%) participants had single visits. The total number of visits for participants with more than one assessment was 116,211. Among 25,337 participants with at least 2 visits, 4,718 (18.6%) participants had 3 visits, 3,778 (14.9%) participants had 4 visits, 2,732 (10.8%) participants had 5 visits, 2,170 (8.6%) participants had 6 visits, and 1,587 (6.3%) participants had 7 visits. The remaining participants had the baseline assessment and at least 7 years of follow-up evaluations.

Our prognostic framework was build using easily collected, cost-effective, and non-invasive cognitive/functional tests, i.e., Functional Activities Questionnaire (FAQ) measuring social functioning and activities of daily living (Pfeffer, Kurosaki, Harrah Jr, Chance, & Filos, 1982); Logical Memory IIA Delayed (LOGIMEM) assessing auditory episodic memory for a short story (Abikoff et al., 1987); Mini Mental State Examination (MMSE) evaluating several mental abilities, including short and long-term memory, attention span, concentration, language and communication skills (Cockrell & Folstein, 2002); Digital Span Forward (DIGIF) and Digital Span Backward (DIGIB) measuring auditory attention span (Schofield & Ashman, 1986); WAIS-R Digit Symbol (WAIS) determining psychomotor speed (Silverstein, 1982), and Boston Naming Test (BOSTON) assessing the ability to name objects (Mack, Freed, Williams, & Henderson, 1992). All selected assessments capture domains sensitive to cognitive aging and the early stages of dementia. After removing patient records with missing values for selected input features and including only individuals with MCI diagnosis, two prognostic models were developed using patient data with 3 (n = 768) (Model 1) and 4 (n = 409) visits (Model 2), separated by one year. Accordingly, in Model 1, disease prognosis at Year 2 (Y2) was determined based on cognitive scores at BL and Year 1 (Y1) and cognitive trajectory classes derived using cognitive scores from the BL and Y1 visits ('cognitive trajectories & cognitive scores' models). In Model 2, disease prognosis at Year 3 (Y3) was determined based on cognitive scores at BL, Y1, and Y2 and cognitive trajectory classes established using the BL, Y1, and Y2 cognitive data ('cognitive trajectories & cognitive scores' models). All participants at BL and Y1 in Model 1 and BL, Y1, and Y2 in Model 2 had MCI diagnosis. Among 768 MCI participants with the record of 3 visits (Model 1), 112 converted to dementia at Y2. Out of 409 MCI participants with 4 visits (Model 2), 65 converted to dementia at Y3. Note that after initial data cleaning, the sample size of MCI individuals with 5 visits or more ($n < 100$) was limited and insufficient to obtain unbiased performance estimates for prognostic models.

2.2 Development of prognostic models

The development of the hybrid prognostic framework comprised two key stages. In Stage 1, we grouped longitudinal trajectories for each considered cognitive/functional assessment using an unsupervised ML approach. Three distinct strategies for Stage 1 have been introduced, (1) the implementation of the longitudinal data partitioning method allowing for clustering trajectories based on their shapes; (2) validation of the optimal number of clusters using three different Clustering Validity Indices (CVIs), and (3) application of the fusion-based methods for combining CVIs into the fused normalized CVI scores, averaged for each cluster partition to determine the final number of trajectory classes for each type of clinical scores. Such generated trajectory classes, characterizing the similar symptomatic progression of subjects (e.g., stable vs. decline), were then fed to Stage 2 as additional input variables when building prognostic Models 1 and 2. The performance of these models was later compared to the performance of classifiers developed using only cognitive scores ('cognitive scores' models). The overall procedure for model development is presented in Fig. 1.

2.2.1 Stage 1: Cognitive trajectory modelling using unsupervised learning

Most of partitioning methods applied to the longitudinal data, group together individuals that have close trajectories at given time points (e.g., using Euclidean distance) without adjusting for their shapes (Bhagwat, Viviano, Voineskos, Chakravarty, & Alzheimer's Disease Neuroimaging Initiative, 2018). In effect, trajectories of identical shapes shifted in time are often defined as different and are consequently assigned to different clusters. In the context of our study, the disease progression is more important than the moment at which it occurs and therefore, we classified the cognitive trajectories according to their shapes (rather than classical distances) using the Fréchet distance:

$$d(A,B) = \inf \max_{\alpha,\beta,t\, \in [0,1]} \left(d\left(A\big(\alpha(t)\big), B\big(\beta(t)\big) \right) \right) \quad [1]$$

and the Fréchet mean:

$$m(A,B) = \left(\frac{d\big(A(\alpha(t_1)),B(\beta(t_1))\big)}{2}, \frac{d\big(A(\alpha(t_2)),B(\beta(t_2))\big)}{2}, \dots, \frac{d\big(A(\alpha(t_n)),B(\beta(t_n))\big)}{2} \right) \quad [2]$$

where *A, B* are trajectories, *α(t)* and *β(t)* describe the position on a trajectory at a time point *t ∈ [0, 1].*

We identified the optimal number of clusters based on multiple clustering validation metrics, namely, Silhouette (Rousseeuw, 1987), Calinski-Harabasz (Caliński & Harabasz, 1974), and Dunn (Bezdek & Pal, 1998) cluster validity indices. All selected measures assessed the quality of partitioning by taking into account the compactness of individual points in the same cluster and their separation in the distinct clusters. Initially, each CVI was calculated for several candidate clustering solutions *k ∈ Ω*, where *Ω* is the ordered set of candidate cluster numbers. The preferred number of clusters was then obtained by finding the value of *k* that maximized the function CVI(*k*) over all values from *Ω*.

Since evidence shows that no single CVI can always outperform others (Kryszczuk & Hurley, 2010), we implemented the decision-level fusion of multiple CVIs using four score fusion-based methods presented in (Kryszczuk & Hurley, 2010) and described in Table S1. Before calculating the combined score of Silhouette, Calinski-Harabasz, and Dunn CVIs using these four strategies, we normalized their scores to a common range using the min-max normalization. We then made the final judgment regarding the optimal number of trajectory clusters for each cognitive/functional test by averaging the output of the four fusion-based methods. The produced trajectory classes were used as input variables (with cognitive scores) in a supervised learning task in Stage 2.

2.2.2 Stage 2: Development of a prognostic model for disease progression using supervised learning

The full dataset was randomly divided into two subsets, namely the model development set (90%), used for model training and validation, and the held-out testing set (10%), used to provide an unbiased evaluation of the final model (Barber, 2012). We implemented the nested validation procedure, with the inner loop serving for model/parameters selection and the outer loop assessing the quality of tuned models on the held-out testing set (10 repeats) (Fig. 1). Note that Z-score normalization was applied to transform multi-scaled data inputs into a common range. Given the model development set, the training partition was used to train the model while the validation partition allowed us to fine-tune the model hyperparameters (Barber, 2012). The optimal hyperparameter selection was conducted by applying grid search with 10-fold cross validation (CV). Table S2 contains the hyper-parameter search spaces and a set of optimal hyperparameters used for the best performing Model 1 and Model 2. Since evidence shows that predictive models developed using imbalanced datasets tend to generate biased and inaccurate results, we applied the Synthetic Minority oversampling technique (SMOTE) to oversample the under-represented class labels in the training set (i.e., individuals with dementia) prior to model fitting (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). For improved generalization performance of predictive models, we also performed feature selection. Specifically, the Random Forest-Recursive Feature Elimination (RF-RFE) with the 10-fold cross validation (CV) was applied on the model development set. For better replicability, the 10-fold CV procedure was repeated 10 times with different partitions of the data. We found that a combination of all features achieved the highest performance for both prognostic models (i.e., Model 1 based on the patient data from three time points spanned over three years and Model 2 based on the patient data from four time points spanned over four years) (Fig. S3 and S4) and hence, all input variables were retained for model training. We then developed three prognostic models based on RF, SVM, and LR to predict conversion to dementia in patients with MCI and tested their performance on the held-out testing set (10 times). Furthermore, we built an additional kNN ensemble model combining the outputs of individual classifiers, as previous studies indicated that ensemble methods may contribute to improvements in accuracy and robustness of predictions (Vilalta, Giraud-Carrier, & Brazdil, 2009). A pairwise statistical test was used to compare the performances of classifiers incorporating both ‘cognitive trajectories & cognitive scores’ input with models based only on ‘cognitive scores’ input. Differences with $p < 0.05$ were considered statistically significant.

3 Results

3.1 Cognitive trajectory modelling

Results of the four fusion-based strategies used to determine the optimal number of clusters for each cognitive/functional assessment in Models 1 and 2 are shown in Fig. S1 and Fig. S2 respectively. The optimal number of trajectory clusters identified for FAQ (0-30), LOGIMEM (0-25), MMSE (0-30), DIGIF (0-12), DIGIB (0-12), and BOSTON scale (0-30) was two. The larger score range of WAIS (0-100) scale enabled modelling of disease progression with higher specificity and resulted in three distinctive trajectory classes.

The cognitive trajectories and the derived clusters’ means are shown in Fig. 2 and Fig. 3. Table 1 shows the percentage of participants assigned to each cluster. Given Model 1, the cluster assignment based on the trajectory-templates yielded two clusters for the majority of cognitive tests, namely FAQ, MMSE, LOGIMEM, DIGIF, DIGIB, BOSTON, with Cluster 1 subjects having consistently better mean results on cognitive tests than individuals in Cluster 2. Cluster 2 also exhibited greater rates of decline over the studied period. For WAIS, we identified three distinct groups in the sample, with Cluster 3 characterized by the greatest decline in cognitive performance over time. In Model 2, the proportion of assigned trajectory labels for each test was comparable to Model 1. Again, two distinctive patient groups were identified based on FAQ, MMSE, LOGIMEM, DIGIF, DIGIB, BOSTON test results. Accordingly, Cluster 1 mean trajectory of cognitive performance appeared relatively stable while Cluster 2 subjects showed a pattern of worsening over time. The generated trajectory classes were later used as additional input features when building prognostic Models 1 and 2 for predicting the MCI-to-dementia conversion ('cognitive trajectories & cognitive scores' models).

3.2 Development of a prognostic model for disease progression

To test the effectiveness and robustness of our approach, we applied four different ML techniques, namely, RF, SVM, LR, and kNN ensemble. Two prognostic models were developed i.e., Model 1 based on the patient data from three time points spanned over three years and Model 2 based on the patient data from four time points spanned over four years. In Model 1, disease prognosis at Y2 was determined using cognitive scores at BL and Y1 and cognitive trajectory classes derived using cognitive scores from BL and Y1 ('cognitive trajectories + cognitive scores' models). We compared the performance of these models to ones developed using only cognitive scores at BL and Y1 ('cognitive scores' models). In Model 2, prognostic predictions at Y3 were established based on cognitive scores at BL, Y1, Y2, and cognitive trajectory classes defined using the BL, Y1, and Y2 cognitive data ('cognitive trajectories + cognitive scores' models). The performance metrics of these models were compared to ones derived using only cognitive scores at BL, Y1, and Y2 ('cognitive scores' models).

Our analysis showed that all models incorporating cognitive trajectory classes into their design performed better than models based purely on cognitive scores ($p < 0.05$ in all cases). Table 2 shows the mean (and standard deviation) of all performance metrics calculated on 10 randomly selected held-out testing sets. Given Model 1 with three time points, prediction accuracy of all models including cognitive trajectory classes was 3-4% higher than accuracy of models built using only cognitive scores. With combined 'cognitive trajectories & cognitive scores' input, the ensemble model achieved the best accuracy of 85% (3.1%), followed by performance of RF with 84.6% (4.1%) accuracy. Although LR model achieved the lowest accuracy of 79.3% (3.4%), it offered the highest prognostic specificity of 72.7% (14.3%). With only 'cognitive scores' input, performance of all models degraded, with both RF and ensemble models providing top accuracy of 81.4%. Again, LR achieved the lowest accuracy (75.9% (3.3%)) but highest specificity in distinguishing MCI from dementia cases (74% (15.3%)).

For Model 2 developed using patient data from four timepoints, performance metrics were again consistently higher when the combined 'cognitive trajectories & cognitive scores' input was used. In addition, models including the fourth time point offered enhanced predictive performance when compared to ones based on only three time points (Model 1). With combined 'cognitive trajectories & cognitive scores' input, the RF model showed the best accuracy of 87.5% (6.1%), followed by the ensemble and SVM, with accuracy of 86.8% (3.7%) and 86.5% (3.8%) respectively. With only 'cognitive scores' input, all models yielded lower performance, i.e., accuracy of 81.3% (6.5%) for RF, 84% (4.7%) for the ensemble model, and 83.8% (5.0%) for SVM. Again, the best specificity was offered by LR, i.e., 74.2% (11.0%) for models based on 'cognitive scores' input and 81.1% (15.6%) for models including 'cognitive trajectories & cognitive scores' input. It is worth noting that although the proportion of MCI non-converters that were correctly identified by our prognostic framework was high and ranged between 84-97%, the specificity of our models, i.e., the ability to correctly identify MCI converters, varied hugely between different solutions (with a range of 24-81%).

4 Discussion

In this study, we developed a novel hybrid machine learning framework for the analysis of longitudinal data that can be used for the clinical prognosis of dementia in individuals with MCI. Our approach utilizes the output of unsupervised learning to improve the classification accuracy of supervised models. We assessed the performance of all models in a systematic and comprehensive way at various time points during follow-up and showed that exploiting information of unlabelled data (Model 1/ 2 incorporating both ‘cognitive trajectories & cognitive scores’ input) can deliver significantly stronger predictions compared to models that do not use the cognitive trajectory features (Model 1/ 2 incorporating only ‘cognitive scores’ input) ($p < 0.05$ in all cases). Although we did not identify one model that would consistently achieve the best performance in all scenarios, we demonstrated that RF and kNN ensemble models offered the best accuracy in most cases. Given the Model 1 based on the patient data from three assessment visits, the best accuracy was reported for the ensemble model (85.0%), followed by RF (84.6%). For Model 2 based on the patient data from four assessment visits, the RF model offered the best accuracy (87.5%), followed by the ensemble model (86.8%). The LR model achieved the highest specificity i.e., 72.7% in Model 1 and 81.2% in Model 2.

The effective performance of RF and ensemble models in differentiating between progressive and non-progressive MCI subjects have been shown in previous studies (Chen et al., 2012; Velazquez & Lee, 2021). Velazquez and Lee (2021) developed the balanced RF model to determine which Early Mild Cognitive Impairment patients were at risk of developing Alzheimer’s disease (AD) and to identify which clinical features were most relevant for conversion prediction. They used the combination of demographic, brain volume, and cognitive data and achieved the accuracy of approximately 90%, higher than the accuracy of competing SVM, XGBoost, and LR implementations. They also found that neuropsychological assessments were the most discriminative features. Chen et al. (2012) utilized the Bayesian Network framework with ensemble learning to differentiate between MCI converters and non-converters. Their approach was based on the Magnetic Resonance Imaging (MRI) data and yielded prediction accuracy of 81%, with sensitivity and specificity of 63% and 89% respectively. It is worth highlighting that the majority of presented models were trained using data from a single timepoint. In the context of dementia prognosis, input from a single time point lacks information regarding temporal changes in clinical status, which is necessary for measuring cognitive decline. An example of the ML framework used to analyse the multiple time point data was shown in Grassi et al. (2019). Their ensemble approach applied to the longitudinal data (socio-demographic, clinical, neuropsychological) achieved the sensitivity of 77.7%, and specificity of 79.9% for predicting the 3-year conversion to AD in MCI individuals. The effective cognitive impairment classification was based only on easily clinically derived information. Several other studies focused on the development of models for predicting the MCI-to-dementia conversion and models based on the multimodal data approach demonstrated higher specificity than our framework (Clark et al., 2014; Guo et al., 2017; Minhas, Khanum, Riaz, Alvi, & Khan, 2016; Spasov et al., 2019). For example, Guo et al. (2017) reported the specificity of 84.78% and sensitivity of 85.0%, by incorporating both cognitive scores and structural MRI data into the model. The combination of neuroimaging data and cognitive assessments yielded the sensitivity of 87.5%, and specificity of 92.31% in Minhas et al. (2016). Spasov et al. (2019) achieved the sensitivity of 87.5%, and specificity of 85% by combining the demographic, neuropsychological, structural MRI, and APOe4 data. All these models incorporated expensive, labour-intensive, and not readily accessible biomarkers, while current general clinical practice relies relatively heavily on cognitive and functional assessments. In contrast, our prognostic framework effectively utilizes information from widely available, cost-effective, and non-invasive dementia markers and hence, can be easily implemented in daily clinical practice. We demonstrated that cognitive/functional tests can reliably and accurately provide prediction of the decline of cognition from MCI to dementia, in particularly when using the patient data from four time points (accuracy of 87.5%, sensitivity of 92.9%, and specificity of 58.3%).

Although our framework was developed using only a relatively small set of cognitive/functional assessments, it still offers sound and high prognostic performance. The fact it utilizes information that is efficient and easily collected and does not depend on procedures that are cost-ineffective, invasive, or not commonly available, such as MRI and PET imaging, cerebrospinal fluid (CSF) collection, and genetic testing ensures minimal subject burden and improved translation to a clinical setting. High predictive power of cognitive/functional assessments in the task of identification of individuals at risk of dementia has been shown in previous studies (Cui et al., 2011; Gupta & Kahali, 2020). Cui et al. (2011) demonstrated that single-modality predictive models based on functional and neuropsychological test outperformed those built using MRI (62%) and CSF (60%) biomarkers, yielding accuracy of 65%. Although an increasing number of studies develops the multimodal approaches for either differentiating between stages of dementia severity or identifying potential predictors for the decline of cognition from MCI to dementia, claiming their superior performance compared to models based only on cognitive/functional assessments, the question about the trade-off between performance and cost-effectiveness or efficiency of the proposed solutions is still questioned and much debated (Bucholc et al., 2019; Fleisher et al., 2008).

Apart from achieving effective prognostic performance based only on a small number of cognitive/functional tests, the main novelty of our ML framework lies in using the output of the rigorously designed unsupervised learning approach as input in a supervised learning task. Our analysis showed that classification models including cognitive trajectory classes generated through unsupervised learning outperformed models based only on clinical scores from multiple visits. To our knowledge, this is the first time that such an approach has been utilized for dementia prognosis/diagnosis. In fact, unsupervised learning has been rarely applied in dementia context (de Langavant et al., 2018; Escudero et al., 2011; Tosto, Monsell, Hawes, Bruno, & Mayeux, 2016). In de Langavant et al. (2018), hierarchical clustering was used to determine the likelihood of dementia using the data from population-based surveys. Escudero el al. (2011) applied the k-Means clustering to identify individuals with mild cognitive impairment (MCI) that are at higher risk of developing dementia while Tosto et al. (2016) used cluster analysis (CA) to model extrapyramidal signs progression in patients with AD. Furthermore, few studies investigated the application of latent class analysis (LCA) for modelling changes in cognitive function (Sukkar, Katz, Zhang, Raunig, & Wyman, 2012; Williams, Storlie, Therneau, Jr, & Hannig, 2020). However, none of these studies attempted to use the output labels of unsupervised learning as input in supervised learning models.

Several limitations warrant mention. First, bias may arise from the degree of accuracy with which cognitive measurements were taken. All tests used in dementia are subject to random measurement error, meaning that change in scores can occur solely due to random fluctuations (Murray et al, 2021). A further issue is bias associated with so-called ‘practice effects’ defined as improvements in cognitive test performance due to repeated measurements using the same test materials (Jabrayilov, Emons, & Sijtsma, 2016). It has been shown that adults with less impaired cognition at baseline may benefit more from practice than those who are more impaired at baseline (Salthouse, 2010). Moreover, several studies suggested that the accuracy of dementia screening assessments may depend on demographic factors including age, gender, education, and ethnicity (Hambleton & Jones, 1993; Schmand et al., 1995). Another limitation of our study is a relatively short follow-up period, limiting more precise modelling of the shape of cognitive trajectories and identifying additional clusters with more subtle deterioration. Third, the identification of cognitive trajectory classes in our study was based on the assumption that they are homogeneous, discrete entities. However, this assumption of within-group homogeneity is not reflected in the heterogeneous pathological nature of neurodegenerative diseases. Finally, the validation of clustering-derived groups of cognitive trajectories is challenging in the absence of ground truth. Therefore, to ensure that our cognitive clusters are biologically meaningful, the external validation using independent datasets is essential. Future efforts should further explore heterogeneity in disease progression, including the analysis of pre-symptomatic and prodromal disease phases (e.g., amnestic MCI and non-amnestic MCI) as well as different dementia types. Given that our procedure led to the performance gain for all considered classifiers, it should be further tested on other longitudinal datasets, in particular, those available for longer time periods and earlier-stage subjects, to provide additional evidence of its accuracy in generalized applications. Furthermore, the application of our methodology to neuroimaging data may help find different patterns of brain atrophy that can be further grouped and used as an additional input when building prognostic classifiers. We also intend to test our framework using the combined cognitive/functional test data and biomarker input, such as brain imaging, to assess the potential prediction gains offered by incorporating additional information.

5 Conclusion

We developed a novel hybrid prognostic framework utilising longitudinal information encoded in efficient, cost-effective, and non-invasive markers to identify subjects with MCI that are at risk for developing dementia. The main novelty of our approach lies in using the output of the rigorously designed unsupervised learning approach as input in supervised ML models. We assessed the performance of our framework in a systematic and comprehensive way for different number of assessment visits and showed that exploiting information generated through unsupervised learning can deliver stronger predictions in a supervised learning task. As the high predictive performance of our prognostic ML-based framework is further confirmed using different longitudinal datasets, it could be incorporated into the clinical decision support system to automate the care pathway for dementia.

Acknowledgements

The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Funding

This work was supported by the Dr George Moore Endowment for Data Science at Ulster University and Alzheimer’s Research UK.

Declaration of Interest

None

Ethics approval

The National Alzheimer's Coordinating Center Uniform Data Set (NACC-UDS) supported by the National Institute on Aging (NIA) (grant U01AG016976) was approved by the University of Washington Institutional Review Board.

Consent to participate

Written informed consent was obtained from all study participants at the Alzheimer's Disease Research Center where they completed their study visits.

Author Contribution

MB and ST were responsible for the study design. MB was involved in the acquisition of data. The data were analysed by MB and ST. All authors were involved in the interpretation of data. MB drafted the article. All authors revised the work critically. All authors approved the final version of the manuscript for publication and agreed to be accountable for all aspects of the work.

Data availability

The data sets generated and analysed during the current study are available through the publicly available National Alzheimer's Coordinating Center UDS database. The current set includes data from the June 2019 NACC data freeze (proposal nr: 1026).

References

Abikoff, H., Alvir, J., Hong, G., Sukoff, R., Orazio, J., Solomon, S., & Saravay, S. (1987). Logical memory subtest of the wechsler memory scale: Age and education norms and alternate-form reliability of two scoring systems. *Journal of Clinical and Experimental Neuropsychology, 9*(4), 435-448.

Bansal, D., Chhikara, R., Khanna, K., & Gupta, P. (2018). Comparative analysis of various machine learning algorithms for detecting dementia. *Procedia Computer Science, 132*, 1497-1502.

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

Barnes, D. E., Covinsky, K. E., Whitmer, R. A., Kuller, L. H., Lopez, O. L., & Yaffe, K. (2010). Dementia risk indices: A framework for identifying individuals with a high dementia risk. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 6*(2), 138.

Bauer, C. M., Cabral, H. J., & Killiany, R. J. (2018). Multimodal discrimination between normal aging, mild cognitive impairment and alzheimer's disease and prediction of cognitive decline. *Diagnostics, 8*(1), 14.

Beekly, D. L., Ramos, E. M., van Belle, G., Deitrich, W., Clark, A. D., Jacka, M. E., & Kukull, W. A. (2004). The national Alzheimer's coordinating center (NACC) database: an Alzheimer disease database. *Alzheimer Disease & Associated Disorders*, 18(4), 270-277.

Beekly, D. L., Ramos, E. M., Lee, W. W., Deitrich, W. D., Jacka, M. E., Wu, J., Hubbard, J.L., Koepsell, T.D., Morris, J.C., & Kukull, W. A. (2007). The national alzheimer's coordinating center (NACC) database: The uniform data set. *Alzheimer Disease & Associated Disorders, 21*(3), 249-258.

Bezdek, J. C., & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28*(3), 301-315.

Bhagwat, N., Viviano, J. D., Voineskos, A. N., Chakravarty, M. M., & Alzheimer's Disease Neuroimaging Initiative. (2018). Modeling and prediction of clinical symptom trajectories in alzheimer's disease using longitudinal data. *PLoS Computational Biology, 14*(9), e1006376.

Bucholc, M., Ding, X., Wang, H., Glass, D.H., Wang, H., Prasad, G., Maguire, L.P., Bjourson, A.J., McClean, P.L., Todd, S., Finn, D.P., & Alzheimer's Disease Neuroimaging Initiative. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert systems with applications*, 130, 157-171.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods, 3*(1), 1-27.

Celebi, M. E., & Aydin, K. (2016). *Unsupervised learning algorithms.* Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chen, R., Young, K., Chao, L. L., Miller, B., Yaffe, K., Weiner, M. W., & Herskovits, E. H. (2012). Prediction of conversion from mild cognitive impairment to alzheimer disease based on bayesian data mining with ensemble learning. *The Neuroradiology Journal, 25*(1), 5-16.

Clark, D. G., Kapur, P., Geldmacher, D. S., Brockington, J. C., Harrell, L., DeRamus, T. P., Blanton, P.D., Lokken, K., Nicholas, A.P., & Marson, D. C. (2014). Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease. *Cortex, 55*, 202-218.

Cockrell, J. R., & Folstein, M. F. (2002). Mini-mental state examination. *Principles and Practice of Geriatric Psychiatry,* 140-141.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., & Jin, J. S. (2011). Identification of conversion from mild cognitive impairment to alzheimer's disease using multivariate predictors. *PloS One, 6*(7), e21896.

de Langavant, L. C., Bayen, E., & Yaffe, K. (2018). Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: Development and validation study. *Journal of Medical Internet Research, 20*(7), e10493.

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., Lituiev D., Copeland T.P., Aboian M.S., & Mari Aparici, C. (2019). A deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain. *Radiology, 290*(2), 456-464.

Escudero, J., Zajicek, J. P., & Ifeachor, E. (2011). Early detection and characterization of alzheimer's disease in clinical scenarios using bioprofile concepts and K-means. Paper presented at the *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 6470-6473.

Fleisher, A. S., Sun, S., Taylor, C., Ward, C. P., Gamst, A. C., Petersen, R. C., Jack, C.R., Aisen, P.S., & Thal, L. J. (2008). Volumetric MRI vs clinical predictors of alzheimer disease in mild cognitive impairment. *Neurology, 70*(3), 191-199.

Ganguli, M., Dodge, H. H., Shen, C., & DeKosky, S. T. (2004). Mild cognitive impairment, amnestic type: an epidemiologic study. *Neurology*, 63(1), 115-121.

Grassi, M., Rouleaux, N., Caldirola, D., Loewenstein, D., Schruers, K., Perna, G., Dumontier, M., & Alzheimer's Disease Neuroimaging Initiative. (2019). A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures. *Frontiers in neurology*, *10*, 756.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., & Alzheimer's Disease Neuroimaging Initiative. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, 65, 167-175.

Guo, S., Lai, C., Wu, C., Cen, G., Weiner, M. W., Aisen, P., Weiner, M., Petersen, R., Jack Jr, C.R., Jagust, W., & Trojanowki, J.Q. (2017). Conversion discriminative analysis on mild cognitive impairment using multiple cortical features from MR images. *Frontiers in Aging Neuroscience, 9*, 146.

Gupta, A., & Kahali, B. (2020). Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests.*Alzheimer's & Dementia: Translational Research & Clinical Interventions, 6*(1), e12049.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, 12(3), 38-47.

Jicha, G. A., Parisi, J. E., Dickson, D. W., Johnson, K., Cha, R., Ivnik, R. J., Tangalos, E.G., Boeve, B.F., Knopman, D.S., Braak, H. & Petersen, R. C. (2006). Neuropathologic outcome of mild cognitive impairment following progression to clinical dementia. *Archives of neurology*, 63(5), 674-681.

Jabrayilov, R., Emons, W. H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied psychological measurement,* 40(8), 559-572.

Jo, T., Nho, K., & Saykin, A. J. (2019). Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in Aging Neuroscience, 11*, 220.

Kryszczuk, K., & Hurley, P. (2010). Estimation of the number of clusters using multiple clustering validity indices. Paper presented at the *International Workshop on Multiple Classifier Systems,* 114-123.

Mack, W. J., Freed, D. M., Williams, B. W., & Henderson, V. W. (1992). Boston naming test: Shortened versions for use in alzheimer's disease. *Journal of Gerontology, 47*(3), P154-P158.

Minhas, S., Khanum, A., Riaz, F., Alvi, A., & Khan, S. A. (2016). A nonparametric approach for mild cognitive impairment to ad conversion prediction: Results on longitudinal data. *IEEE Journal of Biomedical and Health Informatics, 21*(5), 1403-1410.

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 psychiatrica scandinavica, 119(4), 252-265.

Morris, J. C., Weintraub, S., Chui, H. C., Cummings, J., DeCarli, C., Ferris, S., Foster, N. L., Galasko, D., Graff-Radford, N., Peskind, E. R., Beekly, D., Ramos, E. M., & Kukull, W. A. (2006). The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Disease & Associated Disorders*, 20(4), 210-216.

Murman, D. L. (2015). The impact of age on cognition. In Seminars in hearing (Vol. 36, No. 03, pp. 111-121). Thieme Medical Publishers.

Murray, A. L., Vollmer, M., Deary, I. J., Muniz-Terrera, G., & Booth, T. (2021). Assessing individual-level change in dementia research: a review of methodologies. *Alzheimer's Research & Therapy*, 13(1), 1-13.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, 56(3), 303-308.

Pfeffer, R. I., Kurosaki, T. T., Harrah Jr, C. H., Chance, J. M., & Filos, S. (1982). Measurement of functional activities in older adults in the community. *Journal of Gerontology, 37*(3), 323-329.

Prince, M. J., Wimo, A., Guerchet, M. M., Ali, G. C., Wu, Y. T., & Prina, M. (2015). World Alzheimer Report 2015-The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends.

Ritchie, K., Artero, S., & Touchon, J. (2001). Classification criteria for mild cognitive impairment: a population-based validation study. *Neurology*, 56(1), 37-42.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology,* 24(5), 563.

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in aging neuroscience, 9, 329.

Schmand, B., Lindeboom, J., Hooijer, C., & Jonker, C. (1995). Relation between education and dementia: the role of test bias revisited. Journal of Neurology, *Neurosurgery & Psychiatry,* 59(2), 170-174.

Schofield, N. J., & Ashman, A. F. (1986). The relationship between digit span and cognitive processing across ability groups. *Intelligence, 10*(1), 59-73.

Shree, S. B., & Sheshadri, H. S. (2018). Diagnosis of alzheimer's disease using naive bayesian classifier. *Neural Computing and Applications, 29*(1), 123-132.

Silverstein, A. B. (1982). Two-and four-subtest short forms of the wechsler adult intelligence scale-revised. *Journal of Consulting and Clinical Psychology, 50*(3), 415.

Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., & Alzheimer's Disease Neuroimaging Initiative. (2019). A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease.*NeuroImage, 189*, 276-287.

Stirling, J., Chen, T., & Bucholc, M. (2021). Diagnosing alzheimer's disease using a self-organising fuzzy classifier. *Fuzzy logic* (pp. 69-82) Springer.

Sukkar, R., Katz, E., Zhang, Y., Raunig, D., & Wyman, B. T. (2012). Disease progression modeling using hidden markov models. Paper presented at the *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* 2845-2848.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

Tosto, G., Monsell, S. E., Hawes, S. E., Bruno, G., & Mayeux, R. (2016). Progression of extrapyramidal signs in alzheimer's disease: Clinical and neuropathological correlates. *Journal of Alzheimer's Disease, 49*(4), 1085-1093.

Varatharajan, R., Manogaran, G., Priyan, M. K., & Sundarasekar, R. (2018). Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm. *Cluster Computing, 21*(1), 681-690.

Velazquez, M., Lee, Y., & Alzheimer's Disease Neuroimaging Initiative. (2021). Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects. *Plos One, 16*(4), e0244773.

Vilalta, R., Giraud-Carrier, C., & Brazdil, P. (2009). *Data mining and knowledge discovery handbook*. Springer.

Visser, P. J., Kester, A., Jolles, J., & Verhey, F. (2006). Ten-year risk of dementia in subjects with mild cognitive impairment. *Neurology,* 67(7), 1201-1207.

Williams, J. P., Storlie, C. B., Therneau, T. M., Jr, C. R. J., & Hannig, J. (2020). A Bayesian approach to multistate hidden Markov models: application to dementia progression. *Journal of the American Statistical Association,* 115(529), 16-31.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10), 719-731.

Figures

Fig. 1. Overview of the model development procedure.

Fig. 2. The cognitive trajectories and the clusters' means for each considered cognitive/functional assessment in Model 1. Model 1 was based on the patient data from three time points spanned over three years, with trajectory classes derived using clinical scores from baseline (BL) and Year 1 (Y1) visits.

Fig. 3. The cognitive trajectories and the clusters' means for each considered cognitive/functional assessment in Model 2. Model 2 was based on the patient data from four time points spanned over four years, with trajectory classes derived using clinical scores from baseline (BL), Year 1 (Y1), and Year 2 (Y2) visits.

Table 1. Number (percentage) of individuals identified in each data cluster.

	Model 1			**Model 2**		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
FAQ	607 (79%)	161 (21%)	-	288 (70%)	121 (30%)	-
MMSE	568 (74%)	200 (26%)	-	265 (65%)	144 (35%)	-
LOGIMEM	404 (52%)	364 (48%)	-	234 (57%)	175 (43%)	-
DIGIF	453 (59%)	315 (41%)	-	265 (65%)	144 (35%)	-
DIGIB	474 (62%)	294 (38%)	-	230 (56%)	179 (44%)	-
BOSTON	630 (82%)	138 (18%)	-	316 (77%)	93 (23%)	-
WAIS	345 (45%)	247 (32%)	176 (23%)	207 (50%)	109 (27%)	93 (23%)

Table 2. Model performance measures

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
Model 1 (based on 3 time points)				
Cognitive trajectories + cognitive scores				
Ensemble	85.0 (3.1)	95.4 (1.8)	24.7 (12.0)	60.1 (5.4)
RF	84.6 (4.1)	91.4 (3.2)	47.0 (13.6)	69.2 (6.7)
SVM	84.1 (3.1)	94.5 (2.3)	24.1 (11.0)	59.3 (4.9)
LR	79.3 (3.4)	81.2 (4.4)	72.7 (14.3)	77.0 (6.0)
Cognitive scores				
Ensemble	81.4 (4.2)	91.8 (3.7)	26.3 (12.6)	59.1 (7.3)
RF	81.4 (3.4)	88.4 (4.7)	45.8 (15.0)	67.1 (6.7)
SVM	80.7 (4.9)	90.3 (4.4)	29.3 (14.8)	59.8 (8.1)
LR	75.9 (3.3)	76.1 (4.6)	74.0 (15.3)	75.0 (6.5)
Model 2 (based on 4 time points)				
Cognitive trajectories + cognitive scores				
Ensemble	86.8 (3.7)	96.5 (2.2)	29.4 (13.7)	63.0 (6.3)
RF	87.5 (6.1)	92.9 (4.6)	58.3 (19.1)	75.6 (10.3)
SVM	86.5 (3.8)	96.2 (2.0)	29.4 (13.7)	62.8 (6.4)
LR	83.0 (6.0)	83.8 (7.1)	81.1 (15.6)	82.4 (7.8)
Cognitive scores				
Ensemble	84.0 (4.7)	95.5 (3.0)	26.9 (16.0)	61.2 (8.2)
RF	81.3 (6.5)	88.9 (5.7)	43.7 (23.5)	66.3 (12.2)
SVM	83.8 (5.0)	95.7 (3.0)	24.2 (12.8)	60.0 (6.7)
LR	81.8 (5.3)	83.5 (6.4)	74.2 (11.0)	78.8 (5.9)

Abbreviation: RF, Random Forest; SVM, Support Vector Machine; LR, Logistic Regression; AUC: Area Under the Receiver Operating Characteristic (ROC) curve. Standard deviations are shown in brackets.