Cognitive Computation
https://doi.org/10.1007/s12559-023-10169-w

Prediction of Cognitive Test Scores from Variable Length Multimodal Data in Alzheimer's Disease

Ulyana Morar[1] · Harold Martin[1] · Robin P. M.[1] · Walter Izquierdo[1] · Elaheh Zarafshan[1] · Parisa Forouzannezhad[1] · Elona Unger[2] · Mercedes Cabrerizo[1] · Rosie E. Curiel Cid[3] · Monica Rosselli[4] · Armando Barreto[1] · Naphtali Rishe[1] · David E. Vaillancourt[5,6] · Steven T. DeKosky[5] · David Loewenstein[3] · Ranjan Duara[7] · Malek Adjouadi[1]

Received: 13 April 2022 / Accepted: 27 June 2023

Abstract

Alzheimer's disease (AD) is a neurogenerative condition characterized by sharp cognitive decline with no confirmed effective treatment or cure. This makes it critically important to identify the symptoms of Alzheimer's disease in its early stages before significant cognitive deterioration has taken hold and even before any brain morphology and neuropathology are noticeable. In this study, five different multimodal deep neural networks (MDNN), with different architectures, in search of an optimal model for predicting the cognitive test scores for the Mini-Mental State Examination (MMSE) and the modified Alzheimer's Disease Assessment Scale (ADAS-CoG13) over a span of 60 months (5 years). The multimodal data utilized to train and test the proposed models were obtained from the Alzheimer's Disease Neuroimaging Initiative study and includes cerebrospinal fluid (CSF) levels of tau and beta-amyloid, structural measures from magnetic resonance imaging (MRI), functional and metabolic measures from positron emission tomography (PET), and cognitive scores from the neuropsychological tests (Cog). The models developed herein delve into two main issues: (1) application merits of single-task vs. multitask for predicting future cognitive scores and (2) whether time-varying input data are better suited than specific timepoints for optimizing prediction results. This model yields a high of 90.27% (SD = 1.36) prediction accuracy (correlation) at 6 months after the initial visit to a lower 79.91% (SD = 8.84) prediction accuracy at 60 months. The analysis provided is comprehensive as it determines the predictions at all other timepoints and all MDNN models include converters in the CN and MCI groups (CNc, MCIc) and all the unstable groups in the CN and MCI groups (CNun and MCIun) that reverted to CN from MCI and to MCI from AD, so as not to bias the results. The results show that the best performance is achieved by a multimodal combined single-task long short-term memory (LSTM) regressor with an input sequence length of 2 data points (2 visits, 6 months apart) augmented with a pretrained Neural Network Estimator to fill in for the missing values.

Keywords Longitudinal analysis · Deep learning · Multitask · Multimodal · Missing data · LSTM · Cerebrospinal fluid · Magnetic resonance imaging (MRI) · Alzheimer's disease · Alzheimer's Disease Neuroimaging Initiative (ADNI)

✉ Ulyana Morar
ulymorar@gmail.com

[1] Center for Advanced Technology and Education, Department of Electrical and Computer Engineering, Florida International University, 10555 West Flagler St. EC 3900, Miami, FL 33174, USA

[2] College of Pharmacy, Florida A&M University, Tallahassee, FL, USA

[3] Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, Miami, FL, USA

[4] Department of Psychology, Florida Atlantic University, Boca Raton, FL, USA

[5] Department of Neurology, University of Florida, McKnight Brain Inst, Gainesville, FL 32611, USA

[6] Department of Applied Physiology and Kinesiology at the University of Florida, FL, Gainesville, USA

[7] Mount Sinai Medical Center, Wien Center for Alzheimer's Disease & Memory Disorders, Miami Beach, FL, USA

Published online: 19 July 2023

Introduction

There is a hypothesized biological and phenotypical transition in which a person who is cognitively normal (CN) and most likely asymptomatic could have subtle changes in brain structure and neurochemistry years before Alzheimer's disease (AD) develops [1–4]. This stage can last as long as 20 years for some and with little or no symptoms to warrant a doctor's attention and with the prospects for identifying someone in this preclinical stage being extremely difficult. Once symptoms emerge, the disease progresses from a preclinical phase (with underlying biomarker abnormalities) to a prodromal mild cognitive impairment (MCI), and ultimately several stages of dementia. The ability to predict with some accuracy the rate of progression has enormous significance for patients and their families (who must make plans for the future), selection of potential treatments, and for identifying subtypes of slow and fast progressors in observational studies and clinical treatment trials.

Various factors have been identified which may assist in the prediction of the rate of progression, including cognitive measures, different imaging modalities of the brain, such as a magnetic resonance imaging (MRI), which provides structural information, and positron emission tomography (PET), which provides brain metabolic and functional measurements, and both can help detect signs of brain atrophy and the presence of amyloid plaques in the brain triggered by the buildup of Amyloid-β (αβ) peptides, and cerebral spinal fluid (CSF) measures, including levels of amyloid beta protein and tau [5–13]. Commonly used measures to identify cognitive and functional impairment and to measure progression in these domains include the Mini-Mental State Examination (MMSE), along with the Alzheimer's Disease Assessment Scale (ADAS-Cog) test, a comprehensive assessment tool that detects decline in several cognitive domains, the Rey Auditory Verbal Learning Test (RAVLT), the Functional Activities Questionnaire (FAQ), the Everyday Cognition scale (Ecog), and finally, the Clinical Dementia Rating (CDR), a cognitive and functional scaled sensitive to the severity stages of dementia [14–17]. These four different assessment modalities—MRI, PET, CSF, and cognitive scores—are available in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database and have been key for the development of imaging and machine learning algorithms in the AD field.

A lot of the recent machine learning and deep learning work has been involved in either the classification [18, 19] or prediction of AD. In prior work [20], the authors developed a method to predict the MMSE scores of patients 24 months after the baseline score was recorded. The data used originated from the ADNI database and was pruned to 489 patients that possessed the chosen features: genetic biomarkers, ADNI-based neuropsychological tests memory (ADNI-MEM) and executive function (ADNI-EF), the baseline MMSE score, as well as general demographic data such as gender, age, and education level. The method is based on the use of Support Vector Machines (SVMs) with linear and Radial Basis Function (RBF) kernels. The study proved the potential predictive ability of their selected genetic biomarker. Uniquely, the authors [21] designed a study to leverage the efficacy of the PET modality through image analysis to extract and combine features such as low glucose metabolism paired with a high amyloid deposit as a telling sign of decline in cognitive scores. The dataset used was composed of a subset of the ADNI dataset, with 492 subjects, whose scores were assessed after 12 and up to 72 months from baseline.

MRI features were the focus of studies used to predict future cognitive test scores [22–26]. In [23], the authors utilized two key structural MRI biomarkers—segmentation of the hippocampal region and cortical thickness—to predict the MMSE and ADAS scores of 1359 patients from the ADNI database 12 months after their baseline scores. They employed an anatomically partitioned artificial neural network (APANN) composed of multiple hidden layers to encode the latent features of the input data. While the results only show a decent accuracy, it is most likely due to the limited number of features chosen and should instead be interpreted as having selected a good set of biomarkers. The authors [24] focused on what they termed as 2.5D patch extraction method from 3D MRI to reach moderately high accuracy. With 818 patients from the ADNI data, the study used a convolutional neural network (CNN) for feature selection with an extreme learning machine classifier to predict the disease state 3 years after baseline. An improvement on the SVM is reported by incorporating a switching delayed particle swarm optimization to select the SVM parameters [25]. The implementation involved 361 patients from the ADNI database to predict their AD state 36 months from baseline. The authors in [26] used a CNN to select the spatial features from MRI and a recurrent neural network (RNN) to select the longitudinal features with a method that jointly learns and extracts both of these sets of features. Using 830 patients from the ADNI database, the study reports a high classification accuracy between even progressive MCI and stable MCI, which is extremely challenging. The authors of [27] focused their efforts on the use of 3D deep learning algorithms on resting-state functional MRI scans to predict the MMSE scores of 331 participants, who have taken both the MMSE and the CDR tests. They utilized a 3D CNN for classification purposes and achieved high accuracy as well. A recent study [28] implemented a weakly supervised learning framework to counteract the effects that the small sample sizes can have on typical machine learning algorithms, to demonstrate the effectiveness of this approach with a CNN-based single-input-multi-output architecture that works on extracting features from MRI scans and predicts the stage for 6,932 MRI scans from multiple databases.

Multimodal approaches, where multiple input modalities are simultaneously used, have been used [29–35] to predict the progression of AD. The authors of [29] set out to predict two variables, MMSE and ADAS scores, as well as classify the AD stage of 186 patients from the ADNI database in a 24-month time interval. The study used three modalities: CSF, from which they manually selected Aβ42, tTau, and pTau features, MRI, and PET, whose features were selected by using multimodal multitask (M3T) learning for the feature selection process. The AD stages were classified by an SVM. The single modality feature selection methods were compared to the M3T method, which yielded the highest accuracy. The multimodal study from [30] included a 48-month running MMSE prediction with steps of 6 months in addition to the previous biomarkers. The multimodal study from [30] included a 48-month running MMSE prediction with steps of 6 months in addition to the previous biomarkers. A distributed M3T learning approach is used over a dataset of 1620 patients for each modality to find the most prevalent features, which are then passed through a gradient boosting technique. The authors achieved a strong correlation overall; as expected, the accuracy went down the further into the future the prediction was. The study could have benefited from making use of the baseline cognitive exam data. In study [31], the authors explored the prediction of future neuropsychological scores of 5 different exams of 1141 patients 24 months from the baseline measurements. Their method of choice was a stochastic gradient boosting of decision trees of multimodal data available on ADNI over an 18-month period. The method used in this study resulted in the highest correlation when compared to other common algorithms such as perceptron or SVMs. The case study [32] made use of measurements from a combination of MRI and PET scans, CSF, and plasma-related biomarkers from 818 patients in the ADNI database. They permutated the different modalities in a multi-source multitask learning architecture to reach moderately high accuracy. The authors of [33] selected 617 patients with MRIs, PETs, CSF measurements, and MMSE scores whose features were fused via their extreme learning machine method. These were then classified to predict the future states of the patients within 3 years of their baseline measurements. The authors in study [34] took a different approach. They chose to produce an extensive multimodal study that harnesses the advantages of artificial intelligent (AI) AD predictors while making the results accessible to doctors and other medical experts with no knowledge of AI and machine learning. Their study takes the following 11 modalities: cognitive scores, MRI, PET, genetics, lab tests, medical history, neuro exams, neuropsychological battery, physical exams, symptoms, and vital signs of 1048 patients from the ADNI database. This is then passed through their two-layer model, which implements the SHapley Additive exPlanations (SHAP) feature attribution framework that turns the black box that machine learning can be into an easily understood in a step-by-step manner.

While these prior studies yielded good results, the use of progressive data can prove to be valuable in terms of the financial cost it would otherwise take to acquire such data. Using numerous and regular time points can certainly improve the performance of algorithms. Studies such as [36–41] have investigated the potential benefits of using long short-term memory (LSTM) neural networks, a type of artificial recurrent neural network popularly used for sequence labeling and prediction, to analyze and forecast AD progression. LSTMs themselves would seem best fitted for trying to predict AD prognosis due to their ability to handle temporally correlated data alongside their clever way of handling the vanishing gradient problem inherent in its predecessors. The authors of [36] set out to investigate LSTM's long-term learning capabilities combined with five biomarkers relating to MRI measurements to predict the course of the disease for 2700 patients taken from the ADNI database for a 30-month duration in time steps of 6 months. Their results were tabulated and compared to previous studies using more prevalent AD prediction algorithms and reported a higher accuracy rate than others, except for the case between MCI and CN, which could be one of the more crucial and challenging predictions to make. The study from [37] chose to use a multimodal amalgamation of CNN, SVM, and LSTM for its temporal and longitudinal advantages to reach a high accuracy rate. Their dataset was made up of the general demographic information, MMSE scores, structural MRI measurements, and the CDR of 416 patients from the Open Access Series of Imaging Studies (OASIS). The method from [38] leverages the LSTM algorithm through a linear discriminant analysis classifier to predict the state of the disease. The study uses the ADNI TADPOLE dataset from which they selected 742 patients with MRI biomarkers to yield a high accuracy rate in addition to showing its consistent capabilities with varying time intervals. The authors from [40] employed a dataset of 1677 from the ADNI TADPOLE challenge covering a period of up to 72 months and used a typical set of biomarkers to be inputted directly into their RNN model. The resulting predictions were compared to those of LSTM, SVM, Linear State Space baseline, and a constant prediction method where their proposed RNN model was reported to have outperformed the other in most cases when predicting ADAS13 cognitive scores. The authors of [41] selected 1526 patients from the ADNI database and implemented a CNN followed by a bidirectional LSTM. They use the patient's cognitive scores as well as additional features from brain scans and CSF samples as inputs to their model and reached a strong accuracy at predicting the AD stage as far out as 84 months from baseline. Other studies [42, 43] use graph convolution networks (GCNs) for early AD diagnosis and a novel sparse regression method to fuse the auxiliary data into the predictor data for the pMCI/sMCI classification.

The aim of this study is to improve upon this foundation with its use of deep learning regression models in an improved time increment of 6 months between nodes over a period of 5 years in addition to limiting the feature selection which would ideally reduce the amount of data needed to be collected in a practical setting. This study develops five different multimodal deep neural networks with different architectures and underlying characteristics, to predict the cognitive test scores of MMSE and ADAS-CoG13 over a span of 60 months (5 years). To the best of our knowledge, no prior study has had predictions every 6 months for up to 5 years using deep learning. Similar to many prior studies, however, we have obtained multimodal biomarker data from Alzheimer's Disease Neuroimaging Initiative. However, in our case, we use longitudinal multimodal data from CSF levels, structural measures from MRI, functional and metabolic measures from PET, and cognitive scores from neuropsychological tests. We present a data augmentation technique and a Bidirectional LSTM Neural Estimator to generate more training and testing samples from the available data and to handle the missing data problem, respectively. The models developed herein delve into two main questions as to ascertain the merits of (1) using single-task learning method that combines 10 individual prediction into 1 output (experiments 2 and 3A) vs. multitask (experiments 1 and 3B) prediction that take advantage of the correlations between modalities from the start through a single deep regressor; and (2) using time-varying input data (Experiment 3A and 3B) vs. the use of a single snapshot of time (Experiments 1 and 2) in order to achieve the highest prediction accuracy of future time points. In this study, single task is used for predicting one time point, while multitask is used for predicting multiple time points simultaneously.

In designing the proposed MDNN models, the following biological features are used as input in the training and testing phases of these models: (1) the apolipoprotein E (APOE) ε4 status at the baseline examination in terms of APOE ε4 positivity, (2) CSF to gauge the measurements of amyloid-β 1–42 peptide (Aβ1-42), total tau (tTau), and tau phosphorylated at the threonine 181 (pTau), and (3) the use of FDG (18-Fluoro-DeoxyGlucose), Pittsburgh Compound-B (PIB), and AV45 in PET imaging so as to assess amyloid positivity when these PET images are registered with the MRI modality.

It is emphasized that (a) the distinct and non-overlapping nature of the testing sets ensures that no data from subjects seen during the training of the model is seen in the testing phase, ensuring a completely independent test set; and (b) the input features of the training and testing sets are standardized (ensuring zero mean and unit standard deviation) using the statistics of the training set, since under practical circumstances, the statistics of the testing set are not known. Furthermore, a grid search approach was used for hyperparameter tunning.

Methods

Study Cohorts

All participating subjects in this study come from the Alzheimer's Diseases Neuroimaging Initiative. ADNI aims to obtain and maintain MRI, PET, CSF, biochemical biomarkers, and neuropsychological tests for the early detection and study of the progression of Alzheimer's disease. All data are publicly available from http://adni.loni.usc.edu/. In all, 1843 ADNI patients between the ages of 55 and 90 were categorized into the following groups with their corresponding sample sizes on the individual clinical diagnosis at baseline and follow-up visits:

(a) 485 CN: are cognitively normal at baseline and remain so thereafter
(b) 71 CNc: progressed from a cognitively normal state to mild impairment
(c) 496 MCI: presented (at baseline) and remained mild cognitively impaired
(d) 338 MCIc: presented as MCI and progressed to AD at some point during the study
(e) 331 AD: subjects diagnosed with Alzheimer's disease
(f) 26 individuals who progressed all the way from CN to AD

The differentiation clinical diagnosis categories are defined by the ADNI protocols and depend on MMSE and CDR cutoffs established by them along with other criteria. Due to the inherent variability of some subjects' performance on the MMSE test at different examinations, some of them demonstrate regression in clinical groups (i.e., from MCI to CN or AD to MCI). Cleaning the dataset of these extraneous points, or outliers, could yield better performance metrics than what could be reasonably expected from the population at large. Therefore, to avoid possibly skewing the test, we also include the unstable CNun (78) and MCIun (18) subjects present therein.

Demographic information includes age, gender, years of education, and APOE ε4 status at the baseline examination. These relevant data are presented in Table 1 to show the variation among the different clinical subgroups along with the relevant statistical tests. The participants' mean age ranged between 69.27 and 76.96 years; the percentage of males

Table 1 Study demographics

		CN	CNc	MCI	MCIc	AD	CN to AD	CNun	MCIun
# of subjects		485	71	496	338	331	26	78	18
Age	Mean	73.16	76.96[a]	73.08[b]	74.01[b]	74.62[a,c]	74.98	69.27	72.19
	SD	(6.07)	(5.70)	(7.50)	(7.11)	(7.87)	(4.31)	(8.01)	(8.14)
Education	Mean	16.49	16.30	15.90[a]	15.91[a]	15.19[a,b,c,d]	15.92	16.65	15.83
	SD	(2.64)	(2.76)	(2.88)	(2.80)	(2.94)	(2.53)	(2.38)	(2.18)
MMSE	Mean	29.10	28.89	27.86[a,b]	26.99[a,b,c]	23.27[a,b,c,d]	29.42[c,d,e]	28.65	27.22
	SD	(1.10)	(1.33)	(1.82)	(1.77)	(2.11)	(0.81)	(1.27)	(2.37)
ADAS13	Mean	9.35	11.61[a]	15.14[a,b]	20.74[a,b,c]	29.85[a,b,c,d]	10.83[c,d,e]	10.45	15.06
	SD	(4.33)	(4.68)	(5.91)	(6.47)	(8.01)	(4.31)	(4.57)	(7.68)
Gender (M/F)	%	44.12/55.88	61.97/38.03	61.09/38.91	58.58/41.42[a]	56.19/43.81[a]	46.15/53.85	52.56/47.44	72.22/27.78
APOE ε4 (0/1,2)	%	71.55/28.45	57.75/42.25[a]	57.86/42.14[a]	35.80/64.20[a,b,c]	31.72/68.28[a,b,c]	53.85/46.15	62.82/37.18	44.44/55.56

Significant significance is at least $P < 0.05$

Significance difference between groups were assessed using one-way analysis of variance and post hoc Tukey tests for continuous variables (age, education, MMSE) and chi-squared test for the gender and APOE ε4 variables

AD Alzheimer disease, *CN* normal control, *CNc* converter CN, *MCI* mild cognitive impairment, *MCIc* converter MCI, *CN to AD* converter from CN to AD, *CNun* CN unstable, *MCIun* MCI unstable, *MMSE* Mini-Mental Examination Test, *APOE ε4* apolipoprotein E

[a] Significantly different from CN

[b] Significantly different from CNc

[c] Significantly different from MCI

[d] Significantly different from MCIc

[e] Significantly different from AD

was statistically larger in MCIc (58.58%) and AD (56.19%) groups than in the CN group. On the other hand, APOE ε4 positivity was more prevalent in the MCIc (64.20%) and AD (68.28%) groups than in any others. The number of education years ranged from 4 to 20, with a mean of 15(± 2) years. The mean baseline MMSE and ADAS13 scores are also identified for each of the groups, showing the expected significant changes from CN to AD.

The proposed longitudinal models covered in this study use the four main modalities of the ADNI dataset:

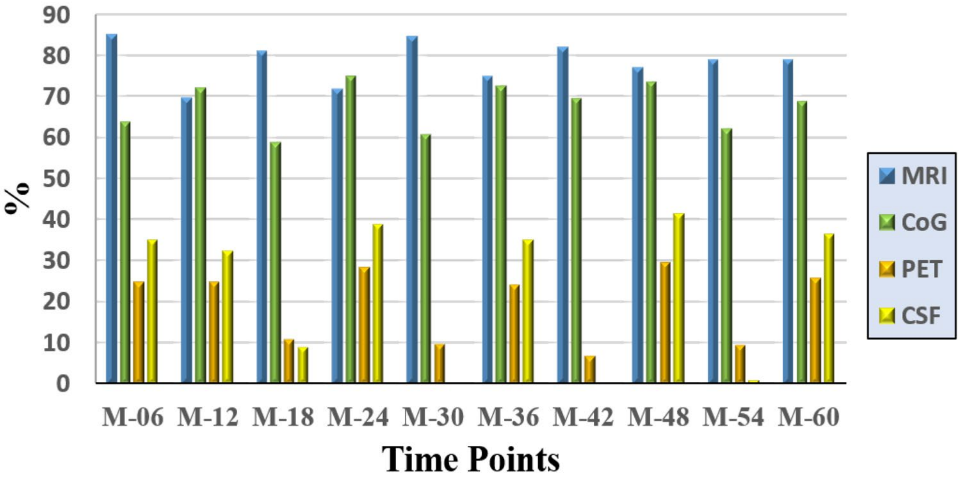
- Neuroimaging measurements from MRIs: ventricular volume, hippocampus volume, whole brain volume, entorhinal cortical thickness, fusiform volume, and the middle temporal gyrus volume. The volumetric measurements are normalized by the intracranial volume (ICV).
- CSF measurements: Aβ1-42, tTau, and pTau.
- PET measurements: FDG, PIB, and AV45.
- Cognitive scores from the: RAVLT, FAQ, Everyday Cognition scales, ADAS13, CDR, and MMSE.

Furthermore, as we are conducting longitudinal studies, we can only include subjects who have at least two separate visits, with at least one more time visit after the initial admission, to validate the predictions. Figure 1 depicts the percentage of available samples for neuroimaging, cognitive, CSF and PET modalities for the different visits at 6 months intervals ($\Delta t = 6$). Understandably, as CSF measurements require a complex and delicate lumbar puncture (spinal tap), the total available samples for it are the least available of all modalities for month 30 (0.1%), month 42 (0.3%), and month 54 (1.1%), while in contrast, MRI data are the most available over the duration of the study.

Data Preparation

Deep learning models like the ones presented in this study require large amounts of data. However, we demonstrate in the next sections that it is still possible to achieve high-performance metrics with a smaller-than-desired number of samples. Recent literature on the longitudinal prognosis of AD tends to only consider baseline biomarker measurements to predict future MMSE scores. This approach does not take advantage of all the other data points available from follow-up visits, greatly reducing the available data for training. In our previous study [44], we implemented a dataset augmentation technique that makes use of the data collected from all available follow-up visits, which go as far as 168 months after the baseline at the time of this study, to generate 8071

Fig. 1 Percent of samples with available data for MRI, CSF, PET, and Cognitive Tests biomarkers for the different time points



distinct samples from the original 1843 patients. This augmentation technique is described in general terms by (1), and an example of its usage is displayed in Fig. 2, as introduced initially in [44].

$$\bigcup_{i=1}^{18}\left[X_{i\Delta t}\right];\left[Y_{(i+1)\Delta t},Y_{(i+2)\Delta t},\dots,Y_{(i+n)\Delta t}\right] \tag{1}$$

$\Delta t = 6,\ n = 10$

where X_i stands for the input features and Y_i stands for the target or predicted values at the i^{th} visit; Δt is the time visit increment (or the minimum discrete time between visits). In our particular case, we link a set of features X at any given time with a corresponding set of targets Y at times ranging from six (6) to sixty (60) months after X is collected.

Through Eq. (1), we generate the set of all possible $[X_i] \rightarrow [Y_{i+\Delta t}, Y_{i+2\Delta t}, \dots, Y_{i+10\Delta t}]$ for every available follow-up visit from month 0 (baseline) to month 108 (i.e., 18×6 with $i = 1, 2, \dots 18$ as in (1)) of every available patient, thereby generating a rich dataset sample from which to train and test the proposed network. Figure 2 shows in practical detail how data from a single subject can generate 18 different samples when using this augmentation technique.

Once the dataset has been augmented using Eq. (1), the full dataset is stratified and randomly divided into training (90%) and testing (10%) sets that are deployed to evaluate and compare the general characteristics of the proposed models. We repeat these steps ten times in accordance with the tenfold cross-validation technique to generate ten distinct and non-overlapping testing sets along with their associated training ensembles.

Data and Code Availability Statement

The clinical data used in this study was obtained from ADNI database. All details pertaining to collection protocols, imaging modalities, and other collected parameters

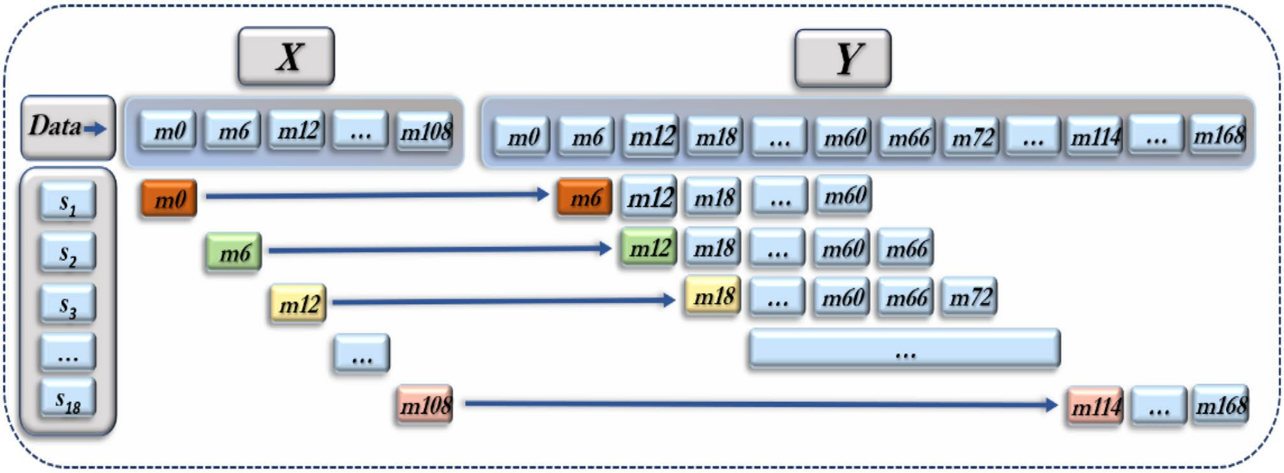


Fig. 2 Dataset augmentation example

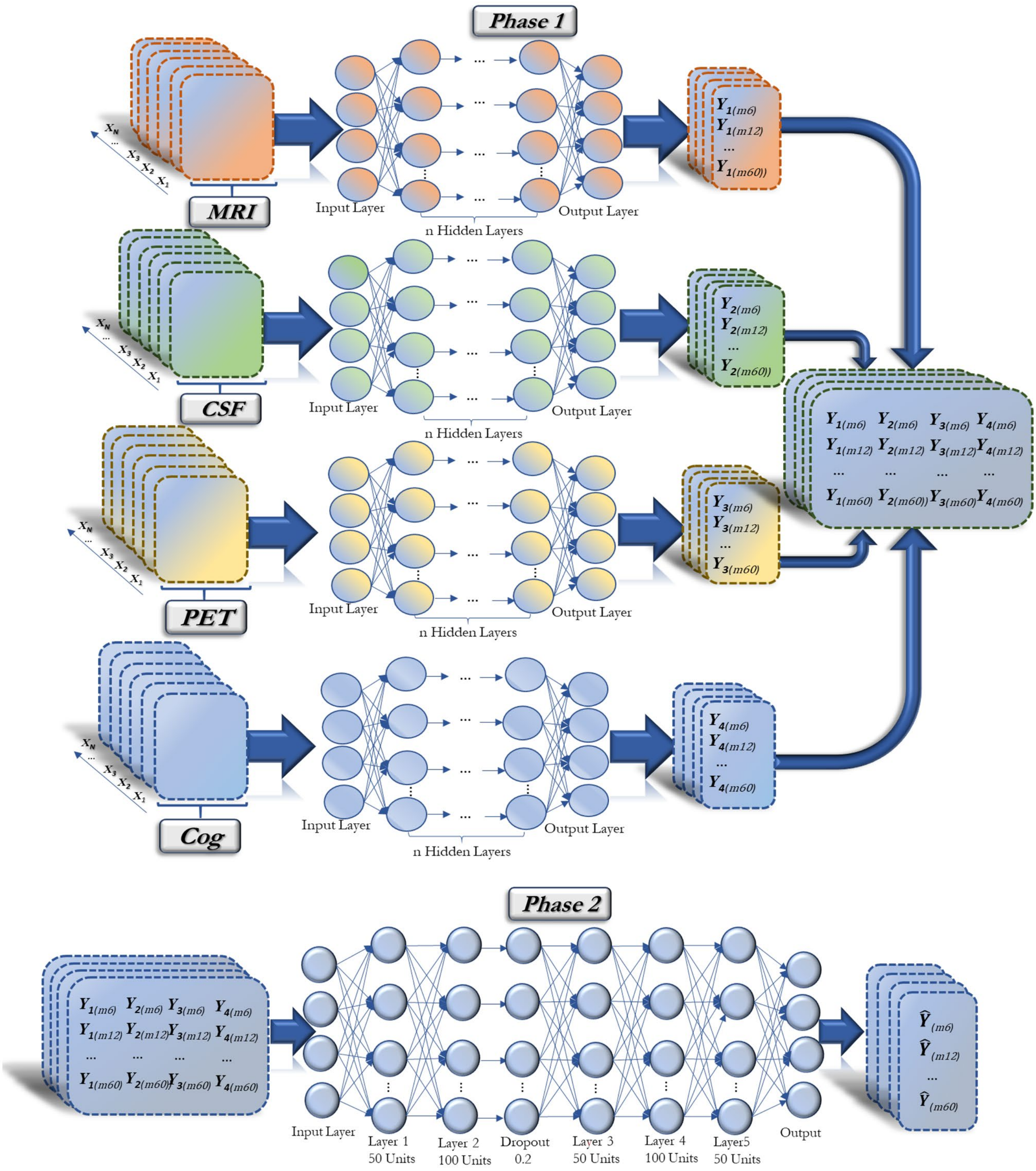

Phase 1
X_N … X_3 X_2 X_1
MRI
CSF
PET
Cog
Input Layer
n Hidden Layers
Output Layer
$Y_{1(m6)}$ $Y_{1(m12)}$ … $Y_{1(m60))}$
$Y_{2(m6)}$ $Y_{2(m12)}$ … $Y_{2(m60))}$
$Y_{3(m6)}$ $Y_{3(m12)}$ … $Y_{3(m60)}$
$Y_{4(m6)}$ $Y_{4(m12)}$ … $Y_{4(m60)}$
$Y_{1(m6)}$ $Y_{2(m6)}$ $Y_{3(m6)}$ $Y_{4(m6)}$
$Y_{1(m12)}$ $Y_{2(m12)}$ $Y_{3(m12)}$ $Y_{4(m12)}$
$Y_{1(m60)}$ $Y_{2(m60))}$ $Y_{3(m60)}$ $Y_{4(m60)}$
Phase 2
Input Layer
Layer 1 50 Units
Layer 2 100 Units
Dropout 0.2
Layer 3 50 Units
Layer 4 100 Units
Layer5 50 Units
Output
$\hat{Y}_{(m6)}$ $\hat{Y}_{(m12)}$ … $\hat{Y}_{(m60)}$

Fig. 3 Combined multitask single-mode regressors (Experiment 1): Neural Network Architecture

can be found at their institutional website (adni.loni.usc.edu). The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The software developed for this study can be made available upon request to the corresponding author of this manuscript.

Table 2 Experimental results for combined multitask single-mode regressors model (Experiment 1)

Time (months):	M06	M12	M18	M24	M30	M36	M42	M48	M54	M60
Correlation, %	85.2	87.58	85.1	85.29	82.02	81.54	78.95	80.06	79.34	76.98
(SD)	(1.61)	(1.28)	(2.13)	(1.85)	(5.21)	(4.2)	(4.92)	(5.18)	(6.09)	(6.82)
RMSE	2.09	2.24	2.48	2.46	2.68	2.66	2.93	2.75	3.28	2.96
(SD)	(0.17)	(0.13)	(0.27)	(0.2)	(0.38)	(0.27)	(0.39)	(0.24)	(0.45)	(0.34)
R^2, %	67.66	72.67	68.61	68.57	63.68	62.18	55.23	59.08	54.77	53.02
(SD)	(4.52)	(3.42)	(4.29)	(2.98)	(8.91)	(7.49)	(11.34)	(10.2)	(9.63)	(14.22)

Experiments

Combined Multitask Single-Mode Regressors (Experiment 1)

In the previous study [44], we combined the four modalities (MRI, CSF, PET, and Cognitive test scores, covered in the "Study Cohorts" section) into a single-input matrix, thereby forcing the regression model to learn their interdependencies and relations to MMSE scores. In contrast, in this experiment, we explored how each of the modalities could separately contribute to the MMSE and how their individual errors can be further decreased by combining their respective single-mode regressor's prediction into the input of a separate regression model. By doing so, as shown in Fig. 3, the new model learns how to optimally reduce the MMSE prediction error of the individual single-mode regressors.

This experiment as illustrated in Fig. 3 consists of two separate phases:

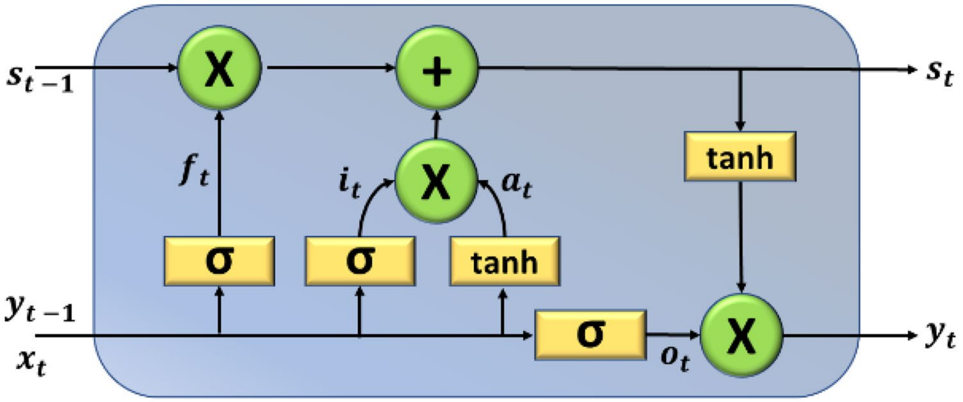
1. Train a deep and fully connected neural network for each modality separately. Similar to the previous experiment, a five-layer network is used with similar training conditions. The first and fourth layers are composed of 50 neurons, the second and third consist of 100, and the last has 10 (one for each of the predicted timepoints). Each layer has Rectified Linear Units (ReLU) activations, L2 regularization of 0.02, and Glorot's weights initialization [45]. The weights are updated using the RMSprop [46] algorithm with a learning rate of 0.01 and 0.8 weight decay.
2. The combined output predictions of Phase 1 of the regressors (predicted MMSE scores from the individual modalities) are combined into a matrix and used as the input for the current stage. These are used to train a new fully connected deep neural network with a total of 6 standard perceptron layers. The first, third, and fifth layers are composed of 50 neurons, the second and fourth of 100, while the last (or output layer) has ten neurons (one for each of the predicted timepoints). Once more, each layer has ReLU activations, L2 regularization of 0.01, and Glorot's weights initialization [45]. RMSprop is again used as the weigh update method, but this time with a learning rate of 0.005 and 0.7 weight decay.

The results of combining the multitask single-mode regressors are shown in Table 2 with the corresponding performance metrics (Correlation, Root-Mean-Squared Error (RMSE), and the Coefficient of Determination (R^2)).

Multimodal LSTM Regressor (Experiment 2)

To take advantage of the longitudinal time-varying nature of the predicted data, we set forth to build a multimodal LSTM regressor. LSTMs are an evolution of RNN that take care of the vanishing gradient problem present in the latter by learning what to remember and what to forget and with the ability to learn and process arbitrarily long sequences. Like RNNs, they are also much better at handling time-varying data than the standard perceptron. LSTMs see and learn the temporal correlations of their neuron's output and inputs.

Fig. 4 LSTM unit structure



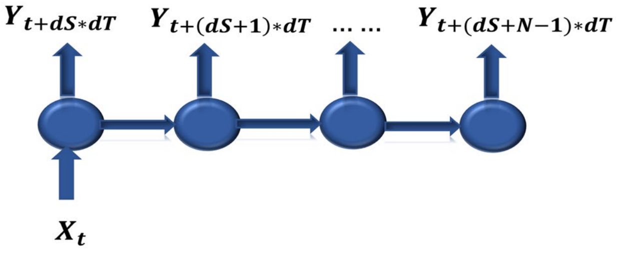


Fig. 5 General one-to-many sequence prediction example for RNNs/LSTMs

Their general principle of operation is described in Fig. 4, along with supporting Eqs. 2 through 8.

$$f_t = S\left(W_f x_t + R_f y_{t-1} + b_f\right) \tag{2}$$

$$i_t = S\left(W_i x_t + R_i y_{t-1} + b_i\right) \tag{3}$$

$$o_t = S\left(W_o x_t + R_o y_{t-1} + b_o\right) \tag{4}$$

$$a_t = \tanh\left(W_c x_t + R_c y_{t-1} + b_c\right) \tag{5}$$

$$s_t = f_t \circ s_{t-1} + i_t \circ a_t \tag{6}$$

$$y_t = o_t \circ \tanh\left(s_t\right) \tag{7}$$

$$S(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

Equation (2) describes the neuron's *Forget Gate*'s behavior, which allows the neuron to choose whether to remember or forget its past internal state when computing its current one. Equation (3) defines the neuron's *Input Gate*; it uses the current input and the neuron's last output to decide whether the current activation should be used in the internal state update. Equation (4) describes whether the neuron will use its current internal state to generate a new output, its *Output Gate*. Equation (5) is the activation function, and Eq. (6) describes how the activation function and the input gate combine with the prior internal state and the forget gate to generate the new internal state. Finally, the output of the independent neurons is expressed by Eq. (7), where its internal state is passed to an activation function (i.e., tanh, ReLU, or others), and the result is allowed to exit the neuron or not based on the *Output Gate*'s state. Equation (8) is the sigmoid activation function. The key advantage of LSTMs over RNNs is in the implementation of these "gates" as they control the information flow through the individual units and their weights and can learn to behave in different ways according to the experiences gained from the training set.

In this experiment, we use the ability of LSTMs to better understand time-varying signals to achieve better MMSE predictions by leveraging the inherent correlation of the scores and the learned features over successive prediction timesteps. The proposed model, as in previous experiments, uses four input modalities (MRI, CSF, PET, and Cognitive test scores) to generate derived time-varying features and produce time-varying MMSE scores. This is a one-to-many model, as it takes a single-input timestep and generates many in return. Figure 5 depicts the general form of these type of models:

Here, *X* defines the input features, *Y* is the predicted output, *t* is time, *dS* is the number of timesteps elapsed between the input features and the first output prediction (in our case one step or six months), *dT* is the length in months of a single timestep (6 months for our experiment), and *N* is the

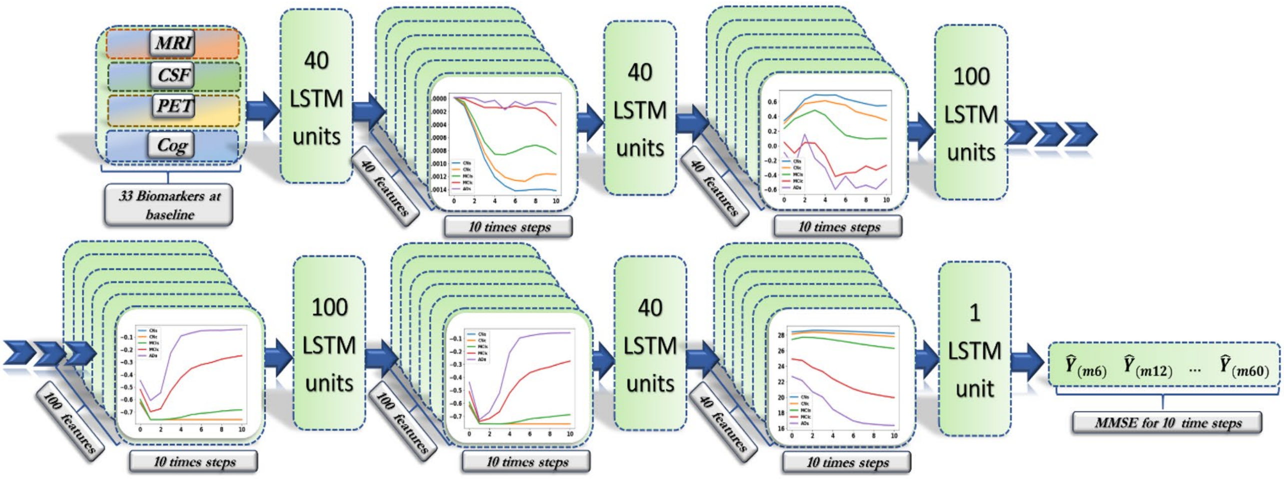


Fig. 6 Multimodal LSTM regressor (Experiment 2). Note: The graphs in front of each layer depict the mean activations of the neuron for each diagnostic group (CN, CNc, MCI, MCIc, AD)

Table 3 Experimental results for multimodal LSTM regressor (Experiment 2)

Time (months):	M06	M12	M18	M24	M30	M36	M42	M48	M54	M60
Correlation, %	84.52	87.27	84.9	85.02	81.58	81.15	78.81	79.96	76.22	75.03
(SD)	(1.66)	(1.28)	(2.4)	(1.99)	(4.29)	(4.92)	(4.82)	(5.9)	(6.95)	(8.77)
RMSE	2.0	2.12	2.37	2.33	2.61	2.53	2.74	2.59	3.2	2.87
(SD)	(0.09)	(0.1)	(0.28)	(0.2)	(0.33)	(0.28)	(0.4)	(0.32)	(0.5)	(0.34)
R^2, %	70.39	75.58	71.3	71.77	65.66	65.62	61.05	63.75	57.15	55.61
(SD)	(2.68)	(1.95)	(3.93)	(3.32)	(7.09)	(7.85)	(7.93)	(9.41)	(9.82)	(13.63)

number of prediction steps defining how far into the future predictions are made.

The network proposed herein, shown in Fig. 6, consists of six layers of LSTM neurons. The first, second, and fifth layers have 40 LSTM units, the third and fourth have 100, and the last has one neuron that produces a different MMSE score prediction for each timestep. Layers 1 through 5 have hyperbolic tangent activations, and the output layer implements a rectified linear activation function. L2 regularization is employed to keep the weights small and avoid weight saturation. The model's weights are randomly initialized, as mentioned earlier, while RMSprop is used to update them at training with a learning rate of 0.01 and 0.8 weight decay.

Table 3 summarizes the main performance metrics (Correlation, RMSE, and the Coefficient of Determination) for the proposed model.

Introduction to Variable Input Length Multimodal LSTM Regressor

In this section, we will explore how the previously proposed LSTM regressor could benefit from an input signal of variable length.

Using a Variable Length Input Sequence

There is an inherent difficulty in trying to estimate or predict a change in a variable from a single set of initial conditions without understanding the model that drives the interrelation between the collected variables. This was the approach tried in the previous section. By contrast, in this section, we take advantage of the temporal correlation of the features in this study and try to produce a model that can learn the inherent correlations between variables over time. The problem at hand can be described visually by the many-to-one model of a single LSTM unit, as shown in Fig. 7.

In this figure, X represents the input feature space, Y is the output prediction, t is time, dT is the time between successive data points, M is the maximum number of collected input points, dS defines the number of timepoints between the last collected X and the predicted Y, and time flows from left to right (i.e., the events on the left happened before the events on the right). The blue circles in this figure represent the same LSTM neuron at different timepoints, and the arrows between them are information (the neuron's internal state) flow over time.

This generalization shows how we can predict any given feature/value from the data collected from multiple prior feature collections. In this section, we will explore how different input sequences—from one to four timepoints in length—can affect the accuracy of the predictions. However, predicting values from input series also has its challenges. In the particular case of the ADNI data, although the visits are discretized in time (i.e., every 6 months), not all patients show up to all the visits. Furthermore, follow-up visits might not be scheduled in the same intervals for every patient and every biomarker modality. Therefore, we must decide not only at which intervals to sample the input sequence but also the length of the sequence to process.

In this experiment, we will consider input sequences with lengths ranging from 1 to 4 visits. Figure 8 provides the number of samples available for the different length input sequences and for the different prediction points. Figure 8a clearly shows that the number of samples is inversely proportional to the length of the input sequence; that is, the longer the sequence, the fewer samples are available; and such is the nature of any longitudinal study, especially of long duration. Figure 8b breaks down the data provided in 8a and displays the number of available samples for each prediction target for different input sequence lengths. For example, the number of samples for input sequences of length 2 (Len2) (that is two consecutive visits 6 months apart, for 6-month window) and length 4 (Len4) (four consecutive visits 6 months apart, or 18-month window) is higher than those of length 1 (Len1, single visit) and length 3 (Len3, 12-month window) for a prediction goal of 6 months in the future (M-06). This

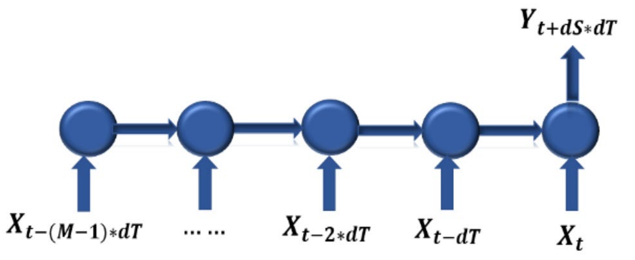


Fig. 7 General many-to-one sequence prediction example for RNNs/LSTMs

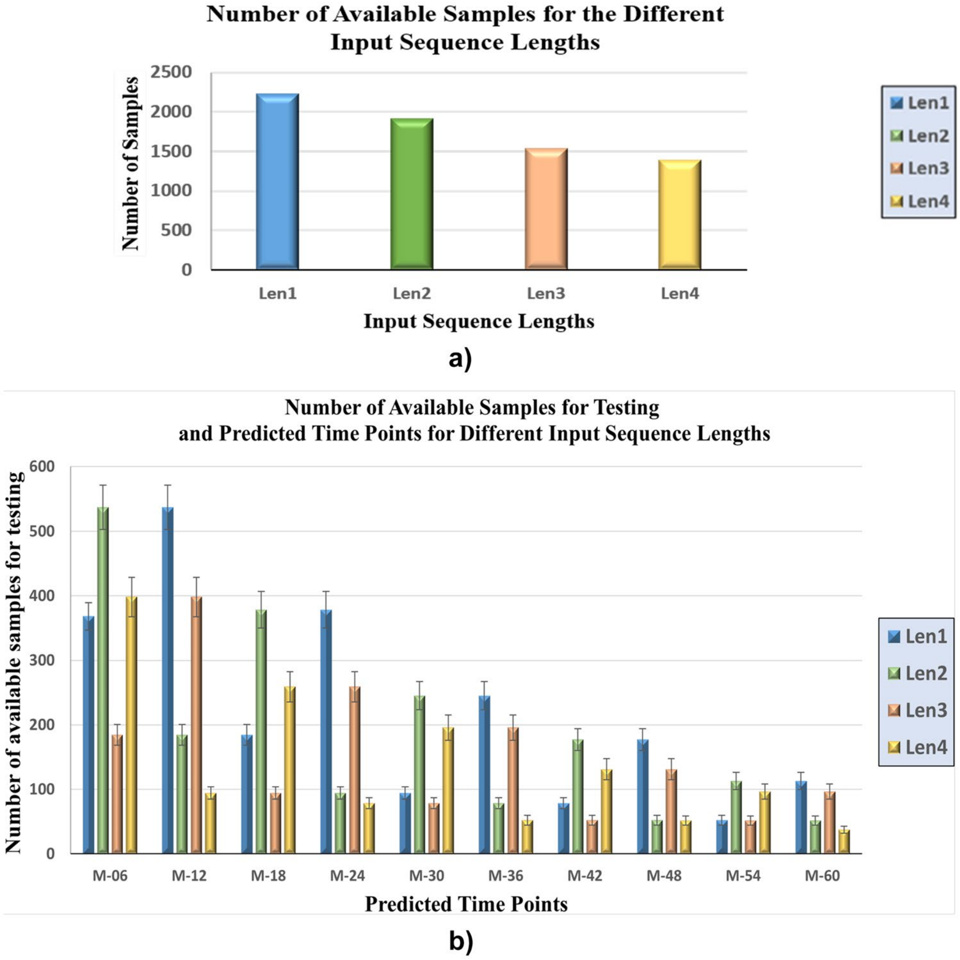


Fig. 8 Number of available samples for different length input sequences: **a** combined; **b** for different prediction points. M – 06 to 60 is the prediction month in the future. Len1 to 4 is a length of the input sequence

is the case because visits are much more frequent in the early phases of the collection and more sparse towards the end or over longer collection windows. In the case of Len1 for M-06, the visit ends up occurring 6 months prior to truth data collection and only cases where two consecutive visits are recorded can be used. However, for Len2, the three visits make up the sequence, baseline (truth collection), 6 month prior to baseline, and 12 month prior to baseline, but only baseline and 2 month prior are required to be valid. This arrangement of Len2 for M-06 is much more common given the 1-year reevaluation period that most patients adhere to as the study progresses.

Handling Missing Values

To overcome the missing value problem, we explore three possible approaches:

(a) Mean Value Imputation (MV): Using the mean value of the feature to fill in any missing values. This is the default procedure after standardization, where missing values (not-a-number (NaN)/null) are replaced by zeros, as zero is the mean of any given standardized vector.

(b) Last Known Value (LKV): Using the last know value for the missing elements of a feature at any given time step is also possible. By using such a method, we create less of a discontinuity between adjacent time points than by using the mean value. However, this creates a stairstep effect.

$X(t) = X(t\text{-}m)$, where $m > 0$ and m is the last know time for which $X(t)$ was not null.

(c) Neural Network Estimator (NNE): Using a neural network is akin to using an enhanced or reinforced interpolation method. This procedure trains a neural network (in our case, a bidirectional LSTM) to predict the missing values of a given feature by using other non-missing ones and prior values of the same and other features.

Neural Network Estimator

We train a bidirectional long short-term memory (BD-LSTM) neural network to predict missing values from the input feature matrix for each of several subgroups of the

Table 4 Results of multiple comparison tests among different filling methods

Methods	**MRI**						**CSF**			**PET**		
	Ventricles	*Hippocampus*	*WholeBrain*	*Entorhinal*	*Fusiform*	*MidTemp*	*ABETA*	*PTAU*	*TAU*	*FDG*	*PIB*	*AV45*
MV vs. LKV	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.001**	**0.047**	**0.047**	**< 0.001**	**0.001**	0.830
MV vs. NNE	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.062	**< 0.001**
LKV vs. NNE	0.644	0.018	0.108	0.003	0.07	**0.003**	**0.005**	**0.005**	**0.005**	**< 0.001**	**0.007**	**< 0.001**
Methods	**COG1**			**COG2**				**COG3**	**COG4**			
	CDRSB	*ADAS13*	*MMSE*	*RAVLT_i*	*RAVLT_l*	*RAVLT_f*	*RAVLT_pf*	*FAQ*	*EcPtMem*	*EcPtLang*	*EcPtVisspat*	*EcPtOrgan*
MV vs. LKV	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.002**	**0.002**	**0.04**	**< 0.001**	**< 0.001**	**< 0.001**	**0.009**	**< 0.001**
MV vs. NNE	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.074	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**0.008**	**0.001**
LKV vs. NNE	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	0.124	**< 0.001**	0.126	0.158	0.321	0.126
Methods			**COG4**									
			EcPtDivatt	*EcPtTotal*	*EcSPMem*	*EcSPLang*	*EcSPVisspat*	*EcSPPlan*	*EcSPOrgan*	*EcSPDivatt*	*EcSPTotal*	
MV vs. LKV			**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	
MV vs. NNE			**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	
LKV vs. NNE			0.126	0.193	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	

Statistical significance with a P-value less than 0.05 is presented in bold face font

Multiple comparison was adjusted using a False Discovery Rate (FDR) of $q < 0.05$

MV Mean Value, *LKN* Last Known Value, *NNE* Neural Network Estimator, for the MRI, CSF, PET, COG1, COG2, COG3, COG4 biomarkers input features. Each subgroup contains a set of features that are always collected at the same time (i.e., CSF biomarkers: if one is collected, all CSF biomarkers are present). The different subgroups are broken down as follows:

MRI ventricular, hippocampus, whole brain, fusiform, and middle temporal gyrus volumes; entorhinal cortical thickness
CSF amyloid-β 1–42 peptide (Aβ1-42), total tau (tTau), and tau phosphorylated at the threonine 181 (pTau)
PET FDG (18-Fluoro-DeoxyGlucose), Pittsburgh Compound-B (PIB), and AV45
COG1 CDR, ADAS13, MMSE
COG2 Rey Auditory Verbal Learning Tests
COG3 Functional Activities Questionnaires
COG4 Everyday Cognition (Ecog) scales

A separate prediction model is trained for each of the subgroups to generate predictions for its values based on the other available features/subgroups. This particular approach is helpful, as it estimates the value of the missing features by using the existing ones, and prior knowledge of the relationship between them is acquired during training. These prediction models are composed of four layers of BD-LSTM units, where layers 1 through 3 have hyperbolic tangent activations and with 30, 20, and 30 neurons, respectively. The fourth and last layers have as many neurons as there are features in the target subgroup and have rectified linear activations.

The separate models are trained using RSMProp with a learning rate of 0.0005 and a weight decay factor of 0.9. We further use L2 regularization of 0.02 to avoid weight saturation and mitigate overfitting, as in previous experiments. Training can run for up to 1000 epochs, but early stopping with the patience of 100 is employed to avoid excessively long training runs. This process is repeated following the tenfold cross-validation approach.

To compare the three aforementioned methods, we use their respective RMSEs (Table S2, Supplementary Section) and show the results of multiple comparisons in Table 4. From Fig. 9, it is evident that the proposed Neural Network Estimator outperforms all other imputation methods. Therefore, in subsequent sections, we will be using the proposed Deep-NN estimator (expressed as NNE in Fig. 9) to impute the missing values.

Variable Input Length Multimodal LSTM Regressor (Experiment 3)

In this experiment, we attempt to predict the MMSE score of patients at multiple future timepoints given the multimodal features collected from anywhere between one and four visits. As both the input and output values are time-dependent, and by the reasons previously given, we deem the use of LSTM neuron appropriate.

We also further explore how multitask learning compares to single-task learning. As such, this experiment is broken down into two sub experiments A and B. Firstly, on experiment A, we combine multiple single-task regressors and produce a Combined Single-Task regressor. In experiment B, we train a multitask regressor to take advantage of the existing interdependencies among biomarker modalities with the aim to achieve better performance metrics.

Combined Multimodal Single-Task LSTM Regressor (Experiment 3A)

The architecture of the proposed model is shown in Fig. 10. It is composed of ten distinct single-task multimodal regressors which are trained to predict an MMSE score at a predetermined time in the future. The output of each regressor is combined into a single output feature vector containing the predictions for each of the ten timesteps in the future after

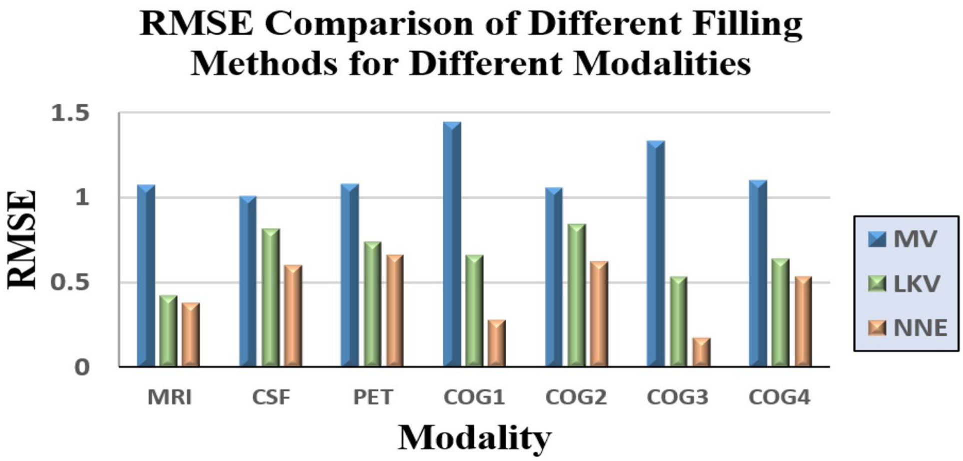


Fig. 9 RMSE comparison of different methods, MV Mean Value, LKN Last Known Value, NNE Neural Network Estimator for MRI, CSF, PET, COG1, COG2, COG3, COG4 modalities

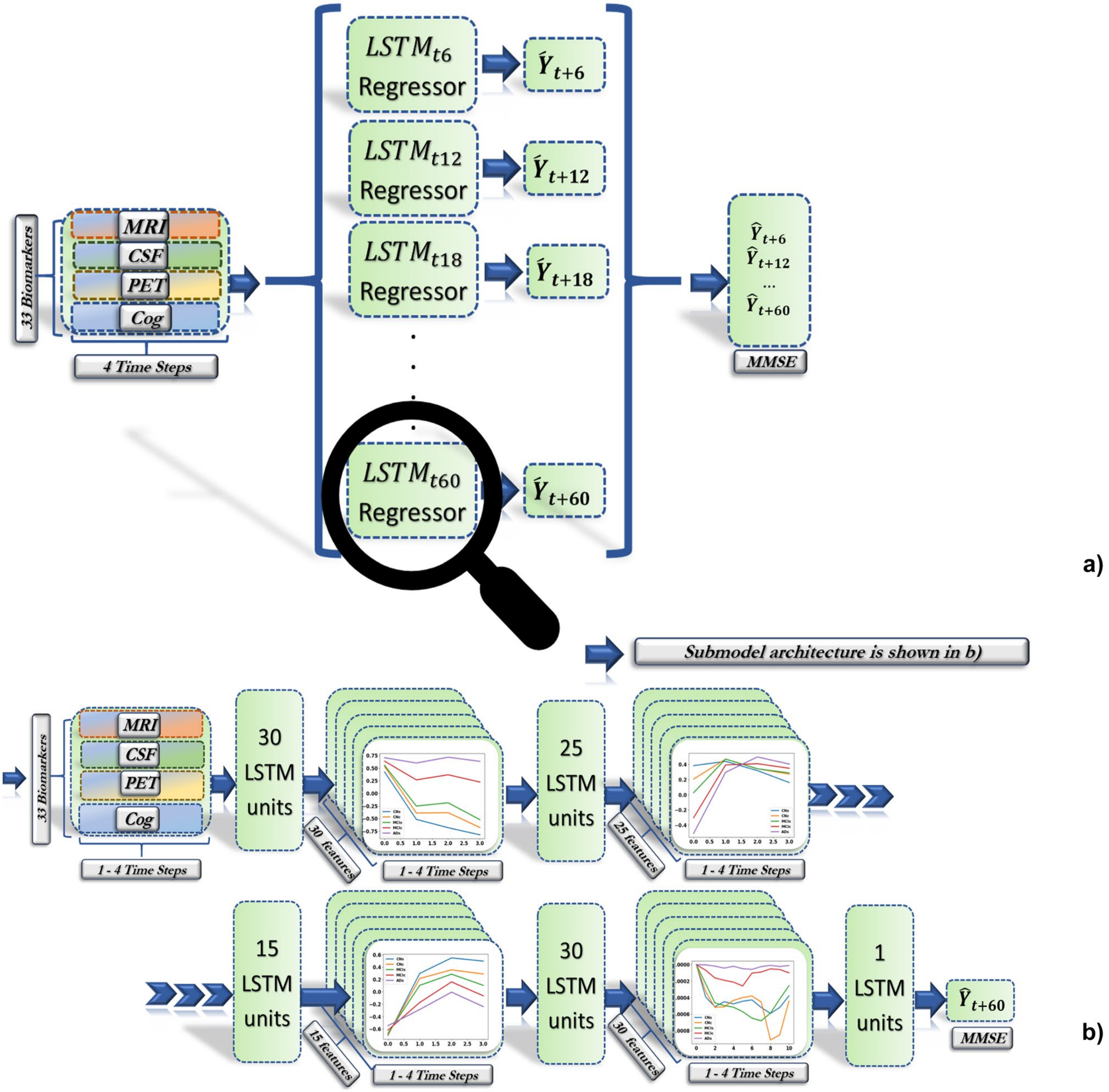


Fig. 10 Combined multimodal single-task LSTM regressor for Experiment 3A. **a** Model architecture. **b** Submodel architecture. Note: The graphs in front of each layer depict the mean activations of the neuron for each diagnostic group (CN, CNc, MCI, MCIc, AD)

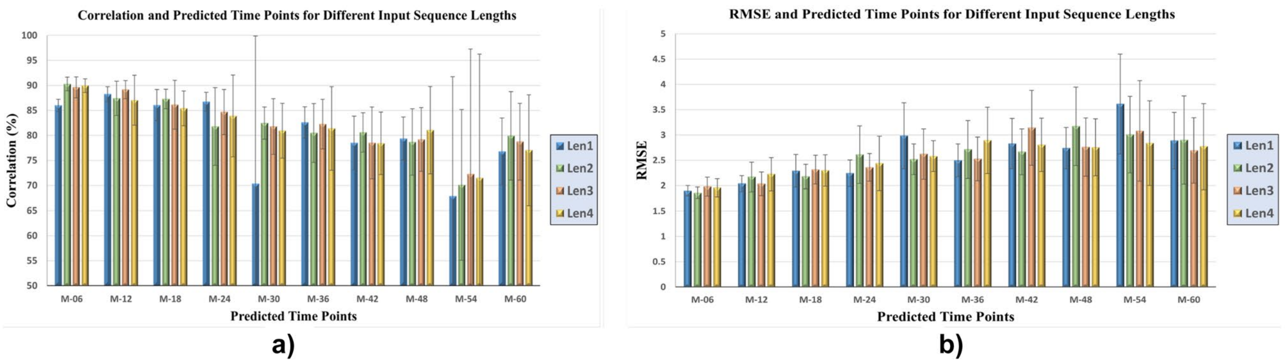


Fig. 11 Performance for different length input sequences (Experiment 3A): **a** Correlation; **b** RMSE

Table 5 Experimental results for combined multimodal single-task LSTM regressor (Experiment 3A)

Time (months):	M06	M12	M18	M24	M30	M36	M42	M48	M54	M60
Correlation, %	90.27	87.42	87.30	81.79	82.48	80.49	80.58	78.68	70.12	79.91
(SD)	(1.36)	(3.42)	(1.94)	(7.78)	(3.21)	(5.87)	(3.95)	(6.66)	(15.05)	(8.84)
RMSE	1.86	2.17	2.18	2.61	2.52	2.71	2.67	3.17	3.01	2.90
(SD)	(0.11)	(0.29)	(0.24)	(0.57)	(0.30)	(0.57)	(0.45)	(0.78)	(0.76)	(0.87)
R^2, %	81.36	75.96	76.06	66.04	67.14	63.60	64.26	59.45	53.66	63.44
(SD)	(2.44)	(6.49)	(3.37)	(12.93)	(5.50)	(10.17)	(6.18)	(9.75)	(20.18)	(13.25)

the last visit in the input sequence. The resulting prediction span covers 60 months.

Each of the individual regression models is composed of five layers of LSTM units, where the first, second, and fourth layers have hyperbolic tangent activations with thirty, twenty-five, and thirty neurons, respectively, while the third and fifth layers have ReLUs and with fifteen and ten neurons, respectively.

The first step in the training process is to pass the input matrix, once standardized, through the Neural Network Estimator proposed in the previous section to fill in for any missing values. Thereafter, the filled-in matrix can be passed on to the model. Training is performed using RSMProp with a learning rate of 0.001 and a weight decay factor of 0.9, while L2 regularization of 0.02 is used to avoid weight saturation and to mitigate overfitting. Just as in the previous experiment, training can run for up to 1000 epochs, but early stopping, with the patience of 100 epochs, is used to avoid excessively long runs. This process is repeated following the tenfold cross-validation approach.

Figure 11 shows the resulting Pearson Correlation and Root-Mean-Squared Error of the trained regressor on the testing set for tenfold cross-validation. From this figure, it is evident that longer sequences do perform better than single-shot predictions (input sequence length of 1), but there is no statistical difference between sequences of length greater than two. Therefore, for the rest of these experiments, we will be reporting the performance metrics associated with input sequences of length two. The statistical comparison of the significance of these differences can be found in Table S1 of the supplementary material.

In Table 5, we summarize the main performance metrics (Correlation, RMSE, and the Coefficient of Determination) for the proposed model for a two-visit input sequence (Len2). The performance obtained for M-54 is viewed as an outlier due to the small sample size available at M-60, where an apparent spike in performance is produced. Table 5 shows a downward slope in the trend of MMSE correlation for M-06, M-18, M-30, M-42, and M-54, which are the months for which most data is available (Figs. 8b (Len2) and 15) and for which we can extract most convincing patterns.

Multimodal Multitask LSTM Regressor (Experiment 3B)

The architecture of the proposed model is shown in Fig. 12. Unlike model in Experiment 3A, this model is composed of a single deep regressor that will try to take advantage of the similarities among the predictive task to learn faster and attain better

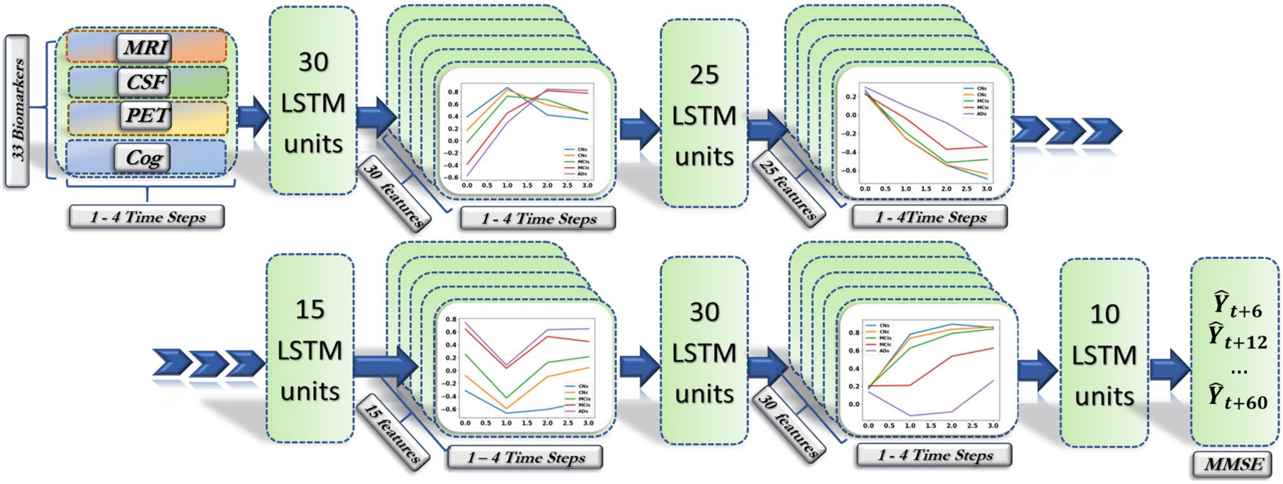


Fig. 12 Multimodal Multitask LSTM Regressor (Experiment 3B). Note: The graphs in front of each layer depict the mean activations of the neuron for each diagnostic group (CN, CNc, MCI, MCIc, AD)

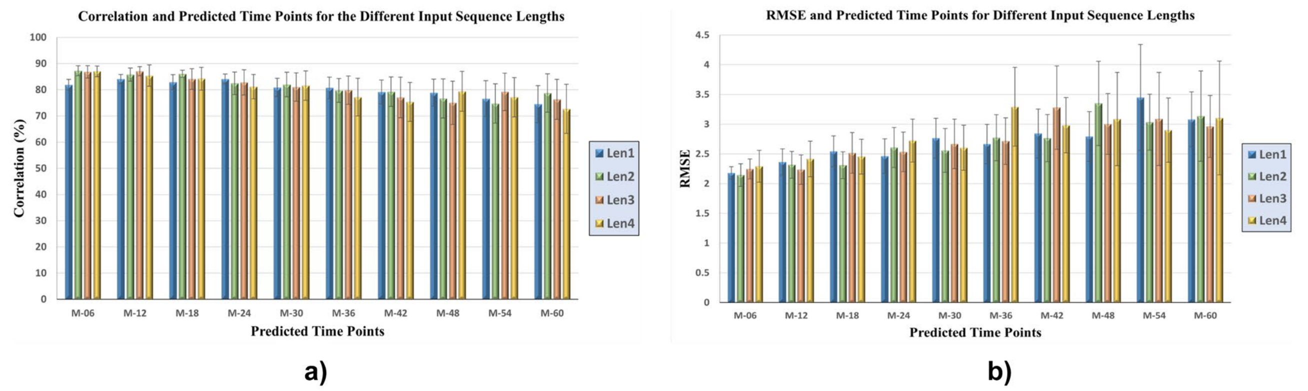


Fig. 13 Performance for different length input sequences (Experiment 3B): **a** Correlation; **b** RMSE

results. The complete model is composed of five layers of LSTM units, where the first, second, and fourth layers have hyperbolic tangent activations and have thirty, twenty-five, and thirty neurons, respectively, while the third and fifth layers have rectified linear activations and fifteen and ten neurons, respectively.

Similar to the previous model, the first step in training the model is to pass the input matrix, once standardized, through the Neural Network Estimator proposed in the previous section to fill in any missing values. Thereafter, the filled-in matrix can be used. Training is also performed using RSMProp with a learning rate of 0.001 and a weight decay factor of 0.9, while L2 regularization of 0.02 is also used to avoid weight saturation and to mitigate overfitting. The process can run for up to 1000 epochs, but, again, early stopping with the patience of 100 epochs is used to avoid excessively long runs. This process is repeated following the tenfold cross-validation approach.

Figure 13 shows the resulting Pearson Correlation and Root-Mean-Squared Error of the trained regressor on the testing set for tenfold cross-validation. In Table 6, we summarize the main performance metrics (Correlation, RMSE, and the Coefficient of Determination) for the proposed model for a two-visit input sequence (Len2).

Results

In this section, we present and compare the individual results of each of the previously described experiments. Starting with Table 7, we present the statistical comparison of each of them to the most encompassing one, the Multimodal Combined Single-Task LSTM regressor (Experiment 3A). Multiple comparisons were corrected with a false discovery rate (FDR) of $q < 0.05$. We provide comparisons in terms of the Pearson Correlation, RMSE, and the Coefficient of Determination. Figure 14 shows these same metrics in plot form for ease of visualization.

From this extensive analysis, it is evident that when predicting the MMSE score for the next 6 months, the model presented in Experiment 3A, using an input sequence of two visits 6 months apart, significantly outperforms all others. However, as the time between the last visit and the intended prediction date increases, the disparity between the models becomes less evident and with no statistical significance. For example, Experiment 3A's correlation was statistically better than Experiment 3B's (*P*-value 0.01), Experiment 2's (*P*-value 0.01), Experiment 1's (*P*-value 0.01), and from the previous work [44] (*P*-value 0.01) when predicting month 6. Furthermore, the RMSE of Experiment 3A is statistically smaller (when predicting the MMSE value for the next 6 months) than that of Experiment 3B (*P*-value 0.01), Experiment 2 (*P*-value 0.049), Experiment 1 (*P*-value 0.02), and [44] (*P*-value 0.001). The RMSE and Coefficient of Determination for the predictions of months 12, 18, and 30 were also statistically better for Experiment 3A than for results from [44].

Furthermore, it is evident from Fig. 14 that, although the range uncertainty grows with time, the performance metrics for Experiment 3A are ahead of those from other experiments

Table 6 Experimental results for multimodal multitask LSTM regressor (Experiment 3B) with two-visit input sequence

Time (months):	M06	M12	M18	M24	M30	M36	M42	M48	M54	M60
Correlation, %	87.22	85.81	86.06	82.47	81.98	79.73	79.24	76.66	74.73	78.72
(SD)	(1.94)	(2.51)	(1.38)	(4.33)	(4.68)	(4.55)	(5.62)	(7.47)	(7.51)	(7.33)
RMSE	2.14	2.32	2.31	2.61	2.56	2.77	2.77	3.35	3.03	3.14
(SD)	(0.19)	(0.23)	(0.23)	(0.34)	(0.37)	(0.39)	(0.40)	(0.71)	(0.47)	(0.76)
R^2, %	75.27	73.00	73.34	67.00	66.63	62.01	60.91	54.72	54.08	57.58
(SD)	(3.37)	(4.23)	(2.69)	(6.75)	(7.58)	(6.51)	(8.50)	(8.28)	(10.46)	(9.17)

Table 7 Statistical comparison of the proposed experiments for MMSE

Correlation (SD), %										
Experiments	**M06**	**M12**	**M18**	**M24**	**M30**	**M36**	**M42**	**M48**	**M54**	**M60**
Combined multimodal single-task LSTM regressor (Experiment 3A)	90.27 (1.36)	87.42 (3.42)	87.30 (1.94)	81.79 (7.78)	82.48 (3.21)	80.49 (5.87)	80.58 (3.95)	78.68 (6.66)	70.12 (15.05)	79.91 (8.84)
Multimodal multitask LSTM regressor (Experiment 3B)	87.22 (1.94)	85.81 (2.51)	86.06 (1.38)	82.47 (4.33)	81.98 (4.68)	79.73 (4.55)	79.24 (5.62)	76.66 (7.47)	74.73 (7.51)	78.72 (7.33)
P-value	**0.01**	0.813	0.595	0.813	0.813	0.813	0.813	0.813	0.813	0.813
Multimodal LSTM regressor (Experiment 2)	84.52 (1.66)	87.27 (1.28)	84.9 (2.4)	85.02 (1.99)	81.58 (4.29)	81.15 (4.92)	78.81 (4.82)	79.96 (5.9)	76.22 (6.95)	75.03 (8.77)
P-value	**0.01**	0.899	0.125	0.578	0.819	0.877	0.752	0.819	0.752	0.578
Combined multitask single-mode regressors (Experiment 1)	85.2 (1.61)	87.58 (1.28)	85.1 (2.13)	85.29 (1.85)	82.02 (5.21)	81.54 (4.2)	78.95 (4.92)	80.06 (5.18)	79.34 (6.09)	76.98 (6.82)
P-value	**0.01**	0.891	0.135	0.613	0.891	0.815	0.71	0.815	0.638	0.71
Multitask multimodal deep regressor [44]	85.3 (1.5)	87.6 (1.2)	85.3 (2.6)	85.6 (1.8)	81.9 (4.7)	81.7 (4.8)	79.4 (4.8)	79.9 (5.2)	78.4 (6.5)	76.3 (7.3)
P-value	**0.01**	0.878	0.345	0.543	0.834	0.818	0.818	0.818	0.666	0.66
RMSE (SD)										
Combined multimodal single-task LSTM regressor (Experiment 3A)	1.86 (0.11)	2.17 (0.29)	2.18 (0.24)	2.61 (0.57)	2.52 (0.30)	2.71 (0.57)	2.67 (0.45)	3.17 (0.78)	3.01 (0.76)	2.90 (0.87)
Multimodal multitask LSTM regressor (Experiment 3B)	2.14 (0.19)	2.32 (0.23)	2.31 (0.23)	2.61 (0.34)	2.56 (0.37)	2.77 (0.39)	2.77 (0.40)	3.35 (0.71)	3.03 (0.47)	3.14 (0.76)
P-value	**0.01**	0.763	0.763	0.986	0.986	0.986	0.986	0.986	0.986	0.986
Multimodal LSTM regressor (Experiment 2)	2.0 (0.09)	2.12 (0.1)	2.37 (0.28)	2.33 (0.2)	2.61 (0.33)	2.53 (0.28)	2.74 (0.4)	2.59 (0.32)	3.2 (0.5)	2.87 (0.34)
P-value	**0.049**	0.769	0.4	0.418	0.761	0.761	0.784	0.25	0.761	0.922
Combined multitask single-mode regressors(Experiment 1)	2.09 (0.17)	2.24 (0.13)	2.48 (0.27)	2.46 (0.2)	2.68 (0.38)	2.66 (0.27)	2.93 (0.39)	2.75 (0.24)	3.28 (0.45)	2.96 (0.34)
P-value	**0.02**	0.633	0.085	0.633	0.573	0.841	0.455	0.437	0.573	0.841
Multitask multimodal deep regressor [44]	2.42 (0.30)	2.52 (0.28)	2.76 (0.36)	2.71 (0.32)	3.00 (0.43)	2.87 (0.38)	3.11 (0.47)	3.05 (0.31)	3.64 (0.40)	3.23 (0.39)
P-value	**0.001**	**0.035**	**0.005**	0.659	**0.035**	0.599	0.077	0.659	0.07	0.421
Coefficient of Determination (SD), %										
Combined multimodal single-task LSTM regressor (Experiment 3A)	81.36 (2.44)	75.96 (6.49)	76.06 (3.37)	66.04 (12.93)	67.14 (5.50)	63.60 (10.17)	64.26 (6.18)	59.45 (9.75)	53.66 (20.18)	63.44 (13.25)
Multimodal multitask LSTM regressor (Experiment 3B)	75.27 (3.37)	73.00 (4.23)	73.34 (2.69)	67.00 (6.75)	66.63 (7.58)	62.01 (6.51)	60.91 (8.50)	54.72 (8.28)	54.08 (10.46)	57.58 (9.17)
P-value	**0.01**	0.532	0.31	0.954	0.954	0.954	0.547	0.532	0.954	0.532
Multimodal LSTM regressor (Experiment 2)	70.39 (2.68)	75.58 (1.95)	71.3 (3.93)	71.77 (3.32)	65.66 (7.09)	65.62 (7.85)	61.05 (7.93)	63.75 (9.41)	57.15 (9.82)	55.61 (13.63)

Table 7 (continued)

Coefficient of Determination (SD), %										
P-value	**0.01**	0.861	0.05	0.522	0.701	0.701	0.547	0.547	0.701	0.522
Combined multitask single-mode regressors(Experiment 1)	67.66 (4.52)	72.67 (3.42)	68.61 (4.29)	68.57 (2.98)	63.68 (8.91)	62.18 (7.49)	55.23 (11.34)	59.08 (10.2)	54.77 (9.63)	53.02 (14.22)
P-value	**0.005**	0.356	**0.005**	0.801	0.52	0.908	0.147	0.935	0.935	0.268
Multitask multimodal deep regressor [44]	56.01 (12.01)	65.06 (7.50)	60.66 (8.75)	61.71 (7.75)	54.12 (12.68)	54.85 (14.41)	47.98 (18.33)	48.25 (18.16)	43.05 (15.92)	43.43 (22.49)
P-value	**0.005**	**0.01**	**0.001**	0.378	**0.027**	0.17	**0.044**	0.154	0.232	0.048

Statistical significance with a P-value less than 0.05 is presented in bold face font

most of the time. This once again shows that LSTMs can take advantage of the inherent correlation present in temporally correlated data to produce better predictions of future sequences and that by leveraging this fact, we are able to better forecast the future progression of MMSE scores.

Figure 15 displays the scatter plots representing the MMSE predictions against the true value along with their corresponding best-fit line for the different diagnosis groups. The predicted values contained within this figure were obtained from the best-performing model (Experiment 3A with two input timesteps). It is evident that as the target time point moves further away from the last observation, the resulting prediction accuracy decreases, along with the number of samples available. This correlation between a decrease in performance and a decrease in the number of available samples is to be expected, as when the number of samples decreases, so does the model's ability to extract and/or learn useful information and features from the limited training scenarios.

Figure 16 shows the evolution of the RMSE metric over the predicted time points for all the individual diagnosis groups, while Fig. 17 breaks down the sample availability of these groups at the different prediction points. From these figures, we can observe how the availability of samples influences the RMSE, and a few interesting patterns emerge: (1) For the CN stable and unstable subgroups, the model's performance remains constant over time. This is to be expected as normal controls' MMSE scores have little variability over time and unstable controls are individuals who might have simply had a bad test day.

A similar pattern is also present in other non-converter subgroups, such as MCI, MCIun, but not in AD. (2) Converter groups' predictive performance decreases over time. MCIc, the biggest converter subgroup, achieves good performance for six (6) and twelve (12) month predictions, but its RMSE rapidly increases afterward, perhaps due to the difficulty in estimating rapid AD progression. CNc also shows the same pattern but to a lesser extent, stabilizing and perhaps decreasing its RMSE after peaking at M30. (3) The number of AD samples available for future prediction rapidly decreases, yielding unstable RMSE scores. From Fig. 17, we can see how the number of AD samples rapidly decreases and becomes almost non-existent after the thirtieth (30th) month timepoint. This yields an unstable performance of the RMSE metric, where, in some instances, it is computed with only two data points.

To further explore the performance of the proposed model on the prediction of the cognitive test scores, we include the prediction metrics for ADAS13-Cog scores in Table 8.

Discussion

In this study, we set out to forecast factors associated with the progression of Alzheimer's disease for a period of 60 months (5 years). The task of prognosticating disease

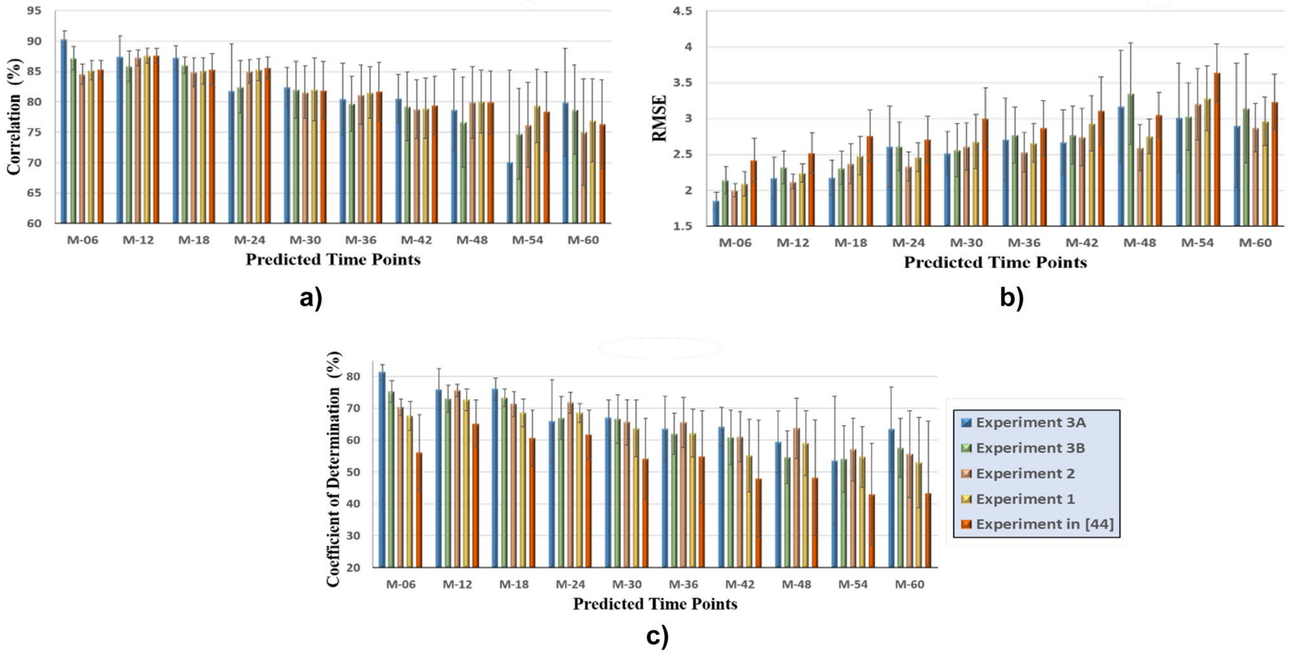


Fig. 14 Performance for different experiments: **a** Correlation; **b** RMSE; **c** Coefficient of Determination

progression has also been studied in the past by multiple researchers. However, we attempt to fill in the gaps of prior research and take better advantage of longitudinal data to generate improved estimates of future cognitive test scores of MMSE and ADAS13.

We start out by proposing a simple multitask multimodal deep regressor in previous study [44]. We use multitask learning, as it has been shown to be beneficial for complex problems where the tasks performed share significant relevant features [30, 31]. While the multimodality allows it to concurrently extract and learn useful information present across the different biomarker modalities, one potential drawback of this approach is that it puts significant pressure on the regressor to learn all the relevant features from the different modalities along with their intercorrelations from a limited set of samples (8071 in this case). Therefore, it might be beneficial to attempt to learn how each modality independently contributes to the cognitive scores of a given test and then build a different regressor to minimize the error of the individual predictions by combining them. This is attempted in Experiment 1, where we see improvements in the testing metrics over results in [44].

In retrospect, these experiments attempt to predict time-dependent signals without using an element capable of truly handling sequential data. Experiment 2 takes on the concept of multitask multimodal learning with the addition of long short-term memory (LSTM) neural networks. This combination yielded a better prediction of RMSE scores than in prior experiments, although not statistically significant. Moreover, Experiment 2, following in the steps of [44] and Experiment 1, attempted to forecast the progression of cognitive test scores from a single timepoint input feature set. Although this is an enticing prospect, we might have been missing out on significant insight that could be extrapolated from a time-varying input sequence.

Therefore, Experiment 3 introduces a variation of Experiment 2 that makes use of variable length input signals. We also cover the missing data problem inherently present in longitudinal studies with multiple input modalities, such as ADNI, and propose a new bidirectional LSTM neural network to fill in for the missing values from prior knowledge of such (i.e., previous data point) and other modalities concurrently available.

Experiment 3 itself is divided into two sub experiments: Experiment 3A combines a series of multimodal single-task regressors, where each regressor is independently tasked to predict a single output timestep, into an output vector containing the ten discrete prediction timepoints; and Experiment 3B makes another pass at multitask learning and combines the regressors of Experiment 3A into a single model. However, similarly to [44] and Experiment 1, the multitask regressor underperforms, and Experiment 3A is shown to yield better performance metrics.

Another significant insight from Experiment 3 is that the best-performing sequence length seems to be two samples long, that is, two visits, 6 months apart.

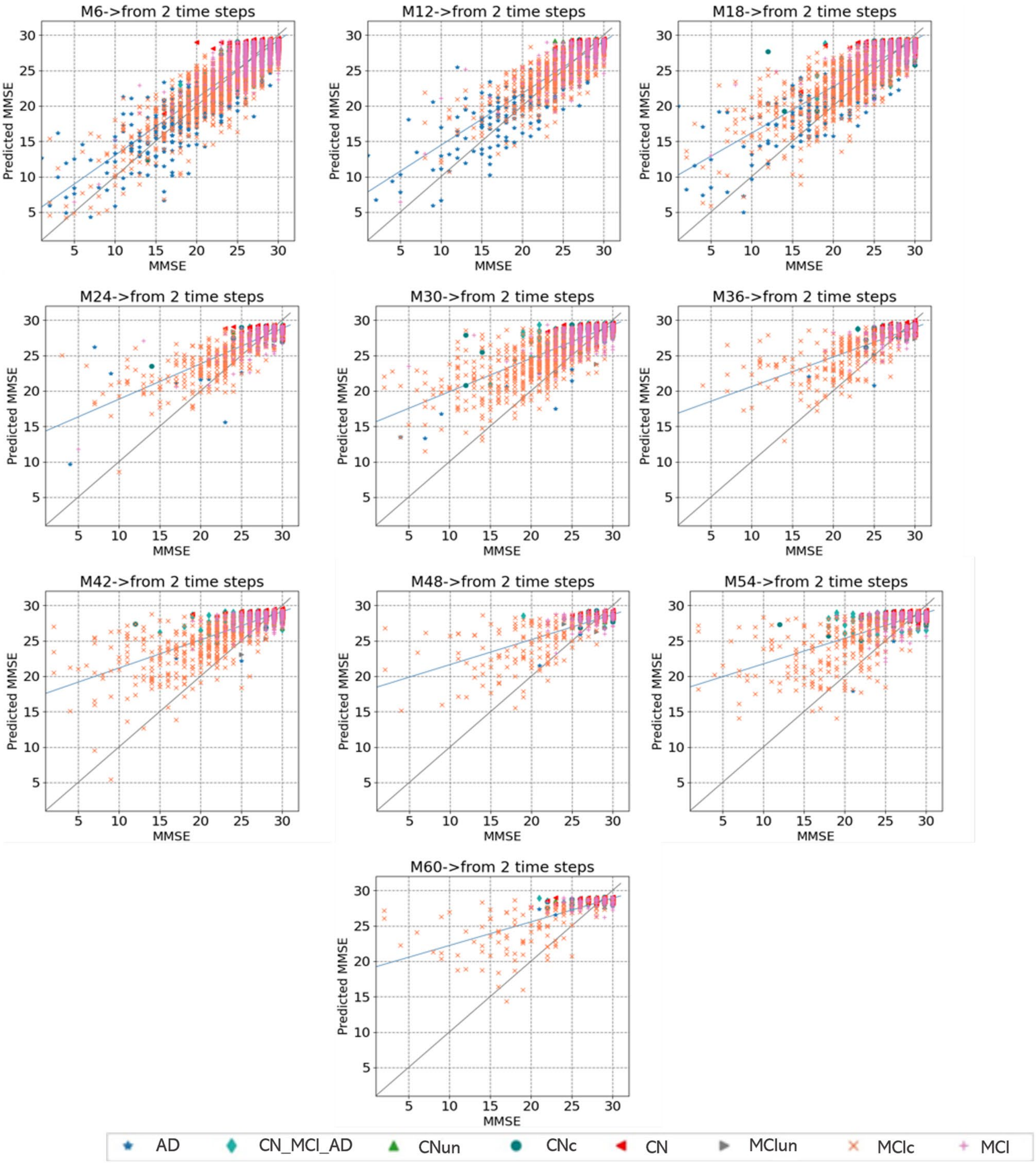


Fig. 15 Scatter plots of predicted and actual MMSE for the different subgroups for Experiment 3A

In the literature, other studies have also attempted to forecast cognitive test scores to varying degrees of success. Tables 9 and 10 show an in-depth performance comparison between the best-performing model from the study, the Multimodal Combined Single-Task LSTM regressor (MCST-LSTM) in Experiment 3A, and the current state-of-art prediction algorithms for MMSE, in terms of Correlation and Root-Mean-Squared Error.

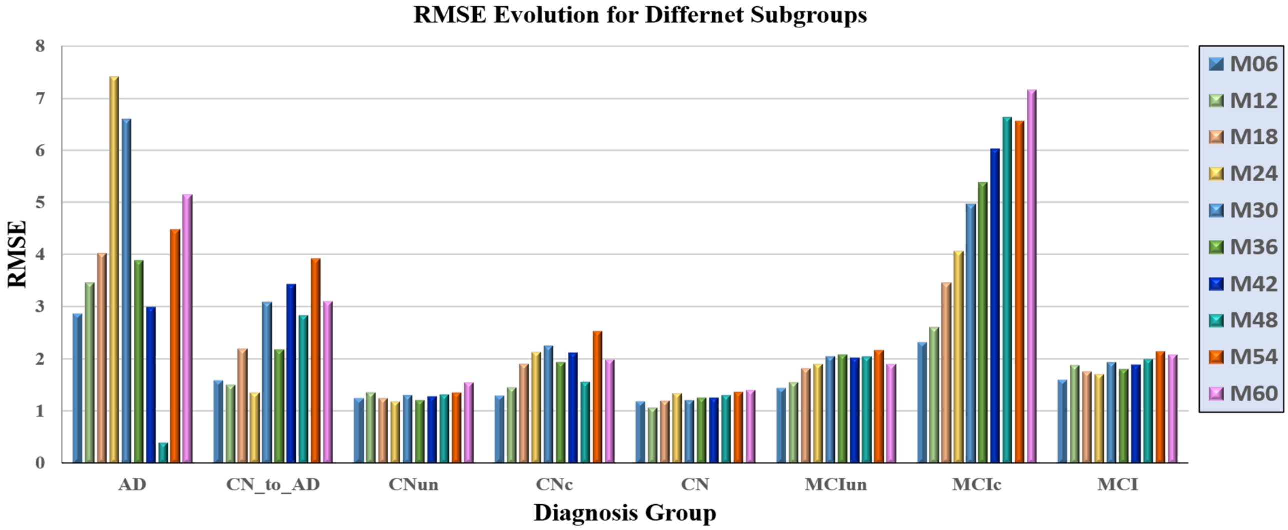


Fig. 16 RMSE evolution for different diagnosis groups for Experiment 3A

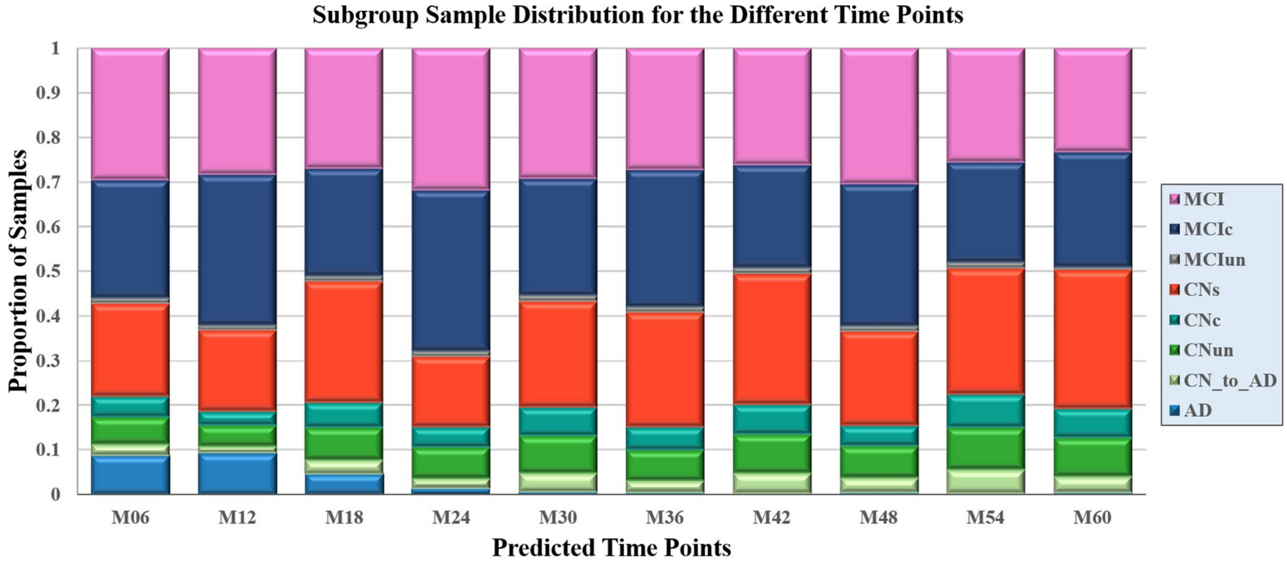


Fig. 17 Subgroup samples distribution for the different time points (Experiment 3A)

Table 8 Experimental results of the best model (Experiment 3A) for ADAS13-CoG

Time (months):	M06	M12	M18	M24	M30	M36	M42	M48	M54	M60
Correlation, %	92.15	91.58	88.95	87.86	85.09	86.51	82.60	81.53	79.32	81.35
(SD)	(0.68)	(1.56)	(1.59)	(3.53)	(3.58)	(4.22)	(3.81)	(7.62)	(6.06)	(6.76)
RMSE	4.74	5.00	5.61	5.84	6.35	6.58	6.93	8.09	7.45	7.27
(SD)	(0.25)	(0.38)	(0.58)	(0.88)	(0.84)	(1.05)	(1.09)	(1.89)	(1.31)	(1.92)
R^2, %	84.80	83.30	78.78	76.56	72.28	74.02	67.49	64.99	61.74	64.96
(SD)	(1.22)	(2.78)	(2.83)	(6.36)	(6.04)	(7.26)	(6.57)	(12.09)	(9.25)	(10.45)

Table 9 Prediction correlation comparison across multiple methods for MMSE, % (SD)

Method	Subjects (samples)	Approach	Modalities	Input sequence length	Time (month)									
					6	12	18	24	30	36	42	48	54	60
[29]	167 (167)	M3T	MRI, FDG-PET, CSF	1	n/a	n/a	n/a	51.1 (2.1)	n/a	n/a	n/a	n/a	n/a	n/a
[30]	1620 (1620)	DMM	MRI, PET, CSF, Cog, Demog, APOE4	1	85.8	79.8	n/a	81.2	n/a	79.0	n/a	75.9	n/a	n/a
[31]	1141 (1141)	GBDT	MRI, PET, CSF, Cog, Demog, APOE4	4	**90.4**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Proposed	1843 (8071)	MCST-LSTM	MRI, PET, CSF, Cog	2	90.27 (1.36)	**87.42 (3.42)**	**87.30 (1.94)**	**81.79 (7.78)**	**82.48 (3.21)**	**80.49 (5.87)**	**80.58 (3.95)**	**78.68 (6.66)**	**70.12 (15.05)**	**79.91 (8.84)**

Table 10 Prediction RMSE comparison across multiple methods for MMSE (SD)

Method	Subjects (samples)	Approach	Modalities	Input sequence length	Time (month)									
					6	12	18	24	30	36	42	48	54	60
[30]	1620 (1620)	DMM	MRI, PET, CSF, Cog, Demog, APOE4	1	1.78 (0.22)	2.24 (0.24)	n/a	2.38 (0.21)	n/a	2.28 (0.22)	n/a	2.19 (0.15)	n/a	n/a
[31]	1141 (1141)	GBDT	MRI, PET, CSF, Cog, Demog, APOE4	4	1.97	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Proposed	1843 (8071)	MCST-LSTM	MRI, PET, CSF, Cog	2	1.86 (0.11)	2.17 (0.29)	2.18 (0.24)	2.61 (0.57)	2.52 (0.30)	2.71 (0.57)	2.67 (0.45)	3.17 (0.78)	3.01 (0.76)	2.90 (0.87)

It is evident from Tables 9 and 10 that the proposed model outperforms the competition across the prediction span of 60 months (represented in bold face font), with the exception of a few months where the performance is tied. For example, we achieved comparable correlation metrics when predicting 6 months into the future, as in [30], but we only use two input time points, in contrast to the four required in [31]. Furthermore, we should point out that we use all the available data to generate the augmented dataset, consisting of 8071 samples, without removing any outliers, which might contribute to the higher variability in our results. Lastly, in Table 10, we report very competitive RMSE scores to those reported in [30] and [31].

Due to the relatively small sample size of the subjects who converted to MCI or AD, we included all subjects in the study. Further investigation of AD progression including only converter groups is still needed to draw a conclusion and compare results with the present study. We did not include demographic variables such as age, sex, or APOE - 4 status in our current study; further studies are still needed to explore the combined effect of including demographics data.

Conclusion

In this study, we explored five different multimodal deep neural networks with different architectures and underlying characteristics, to predict the cognitive test scores of MMSE and ADAS-CoG13 over a span of 60 months (5 years).

The longitudinal multimodal data utilized to train and test the models were extracted from the ADNI study and include CSF levels of tau and beta-amyloid, structural measures from MRI, functional and metabolic measures from PET, and cognitive scores from neuropsychological tests.

We further presented a data augmentation technique to generate more training and testing samples from the available data. We delved into two main issues: (1) by contrasting single ([44], Experiment 2, and Experiment 3A) vs. multitask (Experiment 1 and Experiment 3B) prediction; and (2) determining the suitability of time-varying input data (Experiment 3A and 3B) vs. single snapshot of a time step ([44], Experiment 1, and Experiment 2) to see which of them achieves more accurate predictions of future cognitive scores.

The results show that the best performance is achieved by the Multimodal Combined Single-Task LSTM regressor (Experiment 3A) with an input sequence length of two data points (2 visits, 6 months apart) and a pretrained Neural Network Estimator to fill in for the missing values. This model yields 90.27% (SD = 1.36) correlations for 6 months after the last visit, 87.42% (SD = 3.42) for 12 months, 87.30% (SD = 1.94) for 18 months, 81.79% (SD = 7.78) for 24 months, 82.48% (SD = 3.21) for 30 months, 80.49% (SD = 5.87) for 36 months, 80.58% (SD = 3.95) for 42 months, 78.68% (SD = 6.66) for 48 months, 70.12% (SD = 15.05) for 54 months, and 79.91% (SD = 8.84) for 60 months. These are remarkable findings given the duration of the study and the relatively high accuracy in predicting MMSE and ADAS-CoG13, even for the last time point of 60 months. These results are quite an improvement over previous longitudinal studies, including those from our own research group which considered a 4-year longitudinal study in [30] and a 2-year longitudinal study in [31].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12559-023-10169-w.

Funding This research is supported by the National Science Foundation under grants CNS-1920182, CNS-2018611, and CNS-1551221, and with the National Institutes of Health through NIA/NIH grants 1R01AG055638-01A1, 5R01AG061106-02, 5R01AG047649-05, and the 1P30AG066506-01 with the 1Florida Alzheimer's Disease Research Center (ADRC). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest Mercedes Cabrerizo received support from the National Science Foundation (FIU). Malek Adjouadi received support from the National Science Foundation (FIU), National Institute of Health (UM), and the NIH-1Florida Alzheimer's Disease Research Center (UF), Consulting from UM, and a Speaker Fee from FAMU. David Loewenstein received support from NIH, Statistical Consulting (FIU), and Grand Grounds-Dell Medical Center (Austin Texas). Armando Barreto received support from the National Science Foundation -NSF (FIU) and royalties for his two books from CRC press (Taylor & Francis). David E. Vaillancourt has received research support from NIH, and serves as manager of Neuroimaging Solutions, LLC. Steven T. DeKosky has served as editor (dementia section) and as associate editor for Neurotherapeutics, has served as a consultant on advisory boards, or on data monitoring committees for Acumen Pharmaceuticals, Biogen Pharmaceuticals, Cognition Therapeutics, Prevail Pharmaceuticals, and Vaccinex Pharmaceuticals. Ranjan Duara has received research support from Oregon Health Science University. Authors Ulyana Morar, Walter Izquierdo, Harold Martin, Robin P. M., Parisa Forouzannezhad, and Elaheh Zarafshan received student support from NSF (FIU). Author Elona Unger declares no conflicts of interest with regard to this manuscript.

References

1. Querfurth HW, LaFerla FM. Mechanisms of disease. N Engl J Med. 2010;362(4):329–44.
2. Crous-Bou M, Minguillón C, Gramunt N, Molinuevo JL. Alzheimer's disease prevention: from risk factors to early intervention. Alzheimer's Res Ther. 2017;9(1):1–9.
3. Association A. On the front lines: Primary care physicians and alzheimer's care in america. Alzheimers Dement. 2020;16:64–71.
4. Meek PD, McKeithan EK, Schumock GT. Economic considerations in alzheimer's disease. Pharmacotherapy. 1998;18(2P2):68–73.
5. Morar U, Izquierdo W, Martin H, Forouzannezhad P, Zarafshan E, Unger E, Bursac Z, Cabrerizo M, Barreto A, Vaillancourt DE, DeKosky ST, Loewenstein D, Duara R, Adjouadi M. A study of the longitudinal changes in multiple cerebrospinal fluid and volumetric magnetic resonance imaging biomarkers on converter and non-converter Alzheimer's disease subjects with consideration for their amyloid beta status. Alzheimers Dement (Amst). 2022;14(1):e12258.
6. Fagan AM, Xiong C, Jasielec MS, Bateman RJ, Goate AM, Benzinger TL, et al. Longitudinal change in CSF biomarkers in autosomal-dominant Alzheimer's disease. Sci Transl Med. 2014;6(226):226ra30–226ra30.
7. Loewenstein DA, Curiel RE, DeKosky S, Bauer RM, Rosselli M, Guinjoan SM, et al. Utilizing semantic intrusions to identify amyloid positivity in mild cognitive impairment. Neurology. 2018;91(10):e976–84.
8. Becker JA, Hedden T, Carmasin J, Maye J, Rentz DM, Putcha D, et al. Amyloid-β associated cortical thinning in clinically normal elderly. Ann Neurol. 2011;69(6):1032–42.
9. Gangishetti U, Christina Howell J, Perrin RJ, Louneva N, Watts KD, Kollhoff A, Grossman M, Wolk DA, Shaw LM, Morris JC, Trojanowski JQ, Fagan AM, Arnold SE, Hu WT. Non-beta-amyloid/tau cerebrospinal fluid markers inform staging and progression in Alzheimer's disease. Alzheimer's Res Ther. 2018;10(1):98.
10. Sharma N, Singh AN. Exploring biomarkers for alzheimer's disease. J Clin Diagnostic Res. 2016;10(7):KE01.
11. Bateman RJ, Xiong C, Benzinger TL, Fagan AM, Goate A, Fox NC, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. N Engl J Med. 2012;367:795–804.
12. Pillai JA, Bena J, Bebek G, Bekris LM, Bonner-Jackson A, Kou L, Pai A, Sørensen L, Neilsen M, Rao SM, Chance M, Lamb BT, Leverenz JB; Alzheimer's Disease Neuroimaging Initiative. Inflammatory pathway analytes predicting rapid cognitive decline in MCI stage of Alzheimer's disease. Ann Clin Transl Neurol. 2020;7(7):1225–1239.
13. Fan LY, Tzen KY, Chen YF, Chen TF, Lai YM, Yen RF, et al. The relation between brain amyloid deposition, cortical atrophy, and plasma biomarkers in amnesic mild cognitive impairment and Alzheimer's disease. Front Aging Neurosci. 2018;10:175.
14. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189–98.
15. Kueper JK, Speechley M, Montero-Odasso M. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): modifications and responsiveness in pre-dementia populations. a narrative review. J Alzheimer's Dis. 2018;63(2):423–44.
16. Mohs RC, Knopman D, Petersen RC, Ferris SH, Ernesto C, Grundman M, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. Alzheimer Dis Assoc Disord. 1997;11:13–21.
17. Khan T. Chapter 2-Clinical diagnosis of Alzheimer's disease. Biomarkers in Alzheimer's Disease. 2016;27–48.
18. Helaly HA, Badawy M, Haikal AY. Deep learning approach for early detection of Alzheimer's disease. Cogn Comput. 2022;14(5):1711–27.
19. An N, Ding H, Yang J, Au R, Ang TFA. Deep ensemble learning for Alzheimer's disease classification. J Biomed Inform. 2020;105:103411.
20. Zhu F, Panwar B, Dodge HH, Li H, Hampstead BM, Albin RL, et al. COMPASS: a computational model to predict changes in MMSE scores 24-months after initial assessment of Alzheimer's disease. Sci Rep. 2016;6(1):1–12.
21. Choi H, Jin KH, Initiative ADN, others. Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. Behav Brain Res. 2018;344:103–9.
22. Yan J, Deng C, Luo L, Wang X, Yao X, Shen L, Huang H. Identifying imaging markers for predicting cognitive assessments using wasserstein distances based matrix regression. Front Neurosci. 2019;13:668.
23. Bhagwat N, Pipitone J, Voineskos AN, Chakravarty MM, Initiative ADN, others. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. J Psychiatry Neurosci. 2019;44(4):246–60.
24. Lin W, Tong T, Gao Q, Guo D, Du X, Yang Y, et al. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Front Neurosci. 2018;12:777.
25. Zeng N, Qiu H, Wang Z, Liu W, Zhang H, Li Y. A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease. Neurocomputing. 2018;320:195–202.
26. Cui R, Liu M, Initiative ADN, others. RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Comput Med Imaging Gr. 2019;73:1–10.
27. Duc NT, Ryu S, Qureshi MNI, Choi M, Lee KH, Lee B. 3D-deep learning based automatic diagnosis of Alzheimer's disease with joint MMSE prediction using resting-state fMRI. Neuroinform. 2020;18(1):71–86.
28. Liang S, Gu Y. Computer-aided diagnosis of Alzheimer's disease through weak supervision deep learning framework with attention mechanism. Sensors. 2020;21(1):220.
29. Zhang D, Shen D, Initiative ADN, others. Multi-modal multitask learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage. 2012;59(2):895–907.
30. Tabarestani S, Aghili M, Eslami M, Cabrerizo M, Barreto A, Rishe N, et al. A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study. Neuroimage. 2020;206: 116317.
31. Izquierdo W, Martin H, Cabrerizo M, Barreto A, Andrian J, Rishe N, et al. Robust prediction of cognitive test scores in Alzheimer's patients. In: 2017 IEEE signal processing in medicine and biology symposium (SPMB). 2017. p. 1–7.
32. Nie L, Zhang L, Meng L, Song X, Chang X, Li X. Modeling disease progression via multisource multitask learners: a case study with Alzheimer's disease. IEEE transactions on neural networks and learning systems. 2016;28(7):1508–19.
33. Lin W, Gao Q, Yuan J, Chen Z, Feng C, Chen W, et al. Predicting Alzheimer's disease conversion from mild cognitive impairment using an extreme learning machine-based grading method with multimodal data. Front Aging Neurosci. 2020;12:77.
34. El-Sappagh S, Alonso JM, Islam S, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Sci Rep. 2021;11(1):1–26.
35. Karaman BK, Mormino EC, Sabuncu MR; Alzheimer's Disease Neuroimaging Initiative. Machine learning based multi-modal prediction of future decline toward Alzheimer's disease: an empirical study. PLoS One. 2022;17(11):e0277322.

36. Hong X, Lin R, Yang C, Zeng N, Cai C, Gou J, et al. Predicting Alzheimer's disease using LSTM. IEEE Access. 2019;7:80893–901.
37. Dua M, Makhija D, Manasa PYL, Mishra P. A CNN–RNN–LSTM based amalgamation for Alzheimer's disease detection. J Med Biol Eng. 2020;40(5):688–706.
38. Ghazi MM, Nielsen M, Pai A, Cardoso MJ, Modat M, Ourselin S, et al. Training recurrent neural networks robust to incomplete data: application to Alzheimer's disease progression modeling. Med Image Anal. 2019;53:39–46.
39. Ljubic B, Roychoudhury S, Cao XH, Pavlovski M, Obradovic S, Nair R, Glass L, Obradovic Z. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. Comput Methods Programs Biomed. 2020;197: 105765.
40. Nguyen M, Sun N, Alexander DC, Feng J, Yeo BTT. Modeling Alzheimer's disease progression using deep recurrent neural networks. In: 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI). 2018. p. 1–4.
41. El-Sappagh S, Abuhmed T, Islam SR, Kwak KS. Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing. 2020;412:197–215.
42. Shen HT, Zhu X, Zhang Z, Wang SH, Chen Y, Xu X, et al. Heterogeneous data fusion for predicting mild cognitive impairment conversion. Inf Fusion. 2021;66:54–63.
43. Zhu Y, Ma J, Yuan C, Zhu X. Interpretable learning based dynamic graph convolutional networks for Alzheimer's disease analysis. Inf Fusion. 2022;77:53–61.
44. Morar U, Martin H, Izquierdo W, Forouzannezhad P, Zarafshan E, Curiel RE, et al. A deep-learning approach for the prediction of mini-mental state examination scores in a multimodal longitudinal study. In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI). 2020. p. 761–6.
45. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Teh YW, Titterington M, editors. Proceedings of the thirteenth international conference on artificial intelligence and statistics [Internet]. Chia Laguna Resort, Sardinia, Italy: PMLR; 2010. p. 249–56. (Proceedings of machine learning research; vol. 9). Available from: https://proceedings.mlr.press/v9/glorot10a.html.
46. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on. 2012;14(8):2.