Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Robust hybrid deep learning models for Alzheimer’s progression detection

Tamer Abuhmed [a,*,1], Shaker El-Sappagh [b,c,*,1], Jose M. Alonso [c,*]

[a] Department of Computer Science and Engineering, College of Computing, Sungkyunkwan University, Republic of Korea
[b] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, 13518, Banha, Egypt
[c] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 7 September 2020
Received in revised form 19 November 2020
Accepted 16 December 2020
Available online 25 December 2020

Keywords:
Computer-aided diagnosis
Information fusion
Multimodal multitask learning
Alzheimer’s disease
Alzheimer’s progression
Cognitive scores regression

ABSTRACT

The prevalence of Alzheimer’s disease (AD) in the growing elderly population makes accurately predicting AD progression crucial. Due to AD’s complex etiology and pathogenesis, an effective and medically practical solution is a challenging task. In this paper, we developed and evaluated two novel hybrid deep learning architectures for AD progression detection. These models are based on the fusion of multiple deep bidirectional long short-term memory (BiLSTM) models. The first architecture is an interpretable multitask regression model that predicts seven crucial cognitive scores for the patient 2.5 years after their last observations. The predicted scores are used to build an interpretable clinical decision support system based on a glass-box model. This architecture aims to explore the role of multitasking models in producing more stable, robust, and accurate results. The second architecture is a hybrid model where the deep features extracted from the BiLSTM model are used to train multiple machine learning classifiers. The two architectures were comprehensively evaluated using different time series modalities of 1371 subjects participated in the study of the Alzheimer’s disease neuroimaging initiative (ADNI). The extensive, real-world experimental results over ADNI data help establish the effectiveness and practicality of the proposed deep learning models.

1. Introduction

Alzheimer’s disease (AD) is a neurodegenerative chronic brain disorder that results in progressive memory loss and cognitive impairment. Worldwide, 50 million patients have dementia [1]. By 2030, the number of dementia patients is expected to reach 75.6 million, and by 2050 the number of patients will reach 135.5 million [2]. Dementia is one of the extortionate diseases costing $604 billion worldwide in 2010, AD patients make up 60%–80% of all dementia patients [3]. Caregiving to AD patients is a heavy physical and emotional burden on families and friends. In 2014, caregivers provided 17.9 billion hours of unpaid care to AD patients at a cost of about $217.7 billion in just the U.S. [4]. AD pathology occurs several years before the onset of clinical symptoms, which makes the disease hard to detect in its early stages [5]. Mild cognitive impairment (MCI) is considered an AD prodromal phase, where the progression rate from MCI to AD is at an annual rate of 10%–25% [6]. There are two types of MCI, i.e., stable MCI (sMCI) and progressive MCI (pMCI). pMCI patients convert to AD after a certain period of time, but sMCIs patients do not convert. As AD currently cannot be cured or prevented, early detection of possible pMCI progression, before the occurrence of irreversible brain damages, is of tremendous importance for preventive care, it helps inform personalized medicine and ultimately gives a better quality of life. However, this task is tough due to the high subjectivity and instability in individual patient’s cognitive markers and neuroimaging biomarkers. Machine learning (ML) techniques have been widely used to tackle this challenge [6–10]. Some studies have surveyed the recent ML techniques for AD progression prediction [11–13].

1.1. Complexity of AD progression detection

Applying ML to AD progression modeling is a complex process for several reasons. First, AD data are naturally multimodal data [3,11,14] where each patient has a collection of complementary pieces of different types including magnetic resonance imaging (MRI), positron emission tomography (PET), neuropsychological battery, vital signs, cognitive scores (CSs), medical history, genetics, cerebrospinal fluid (CSF), lab tests, physical examinations, symptoms, neurological examination, and demographics.

[*] Corresponding authors.
E-mail addresses: tamer@skku.edu (T. Abuhmed), shaker.elsappagh@usc.es (S. El-Sappagh), josemaria.alonso.moral@usc.es (J.M. Alonso).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.knosys.2020.106688
0950-7051/© 2020 Elsevier B.V. All rights reserved.

These heterogeneous modalities carry complementary knowledge, which describes the disease status from different viewpoints [15]. For example, CSF and PET are the most appropriate for detecting the accumulation of amyloid-β in the brain years before disease-related structural changes can be detected, on the other hand, an MRI is more sensitive to changes after the first symptoms appear. Moreover, medical history, including age and number of years education; CSs including MMSE, ADAS-cog, CDRSB, FAQ, etc.; and neuropsychological battery features such as RAVLTs have shown to be significant factors in predicting AD progression [10,15–18]. However, most AD studies concentrate on MRI analysis [13,19–21]. Although MRI is critical for AD detection, it is its fusion with other modalities that could provide a holistic picture of a patient's status and could potentially enhance the accuracy of a progression model. This fusion reduces noise by averaging it out over a number of independent data sources. Furthermore, multimodal systems usually yield more comprehensive insights, more precise results, more reliable behaviors, and accordingly more acceptance from the medical community [22–25]. However, how to effectively fuse information from heterogeneous sources remains a challenge. This approach is known as multimodal design. Forouzannezhad et al. [26] used a Gaussian-based model to predict MCI based on demographics, PET, and MRI data, they achieve an accuracy of 78.8%. Second, most AD data are collected over time (i.e. longitudinal or time-series data), which indicates that the disease state at a certain time is dependent on the state at a previous point in time. Therefore, to study the relative longitudinal changes in AD, we must trace pathophysiological changes over a huge number of observations [14]. However, the vast majority of the research does not explore AD's temporal/sequential nature [12]. Accordingly, the AD investigations in previous literature did not analyze any correlation among multimodal signals and how they evolve over time [27]. Traditional time series algorithms have been applied in AD progression, but these techniques have their own limitations [28,29]. Third, to accurately predict AD progression, it is critical to monitor multiple features at the same time in what is called multitask modeling [14]. Various efforts have been made in AD multitask modeling, these efforts generally fall into one of multiple categories, including single-modal single-task classification [6] and regression [7] learning, single-modal multitask regression learning [8], as well as multimodal single-task classification [9] and regression [10] learning. Going a step further and considering several modalities and predicting multiple clinical variables that would be required by a medical expert to assess a patient, this is called multimodal multitask learning, where a task has multiple input sources, and all tasks are related in a chronological sequence [17,30–32]. Modeling AD progression as a multimodal multitask process based on time series data is a challenge [33]. However, recent studies have asserted that this modeling of AD related tasks has delivered promising performance and is more stable than other models [14,30]. Zhang et al. [15] proposed a general method called multimodal multitask (M3T) learning to concurrently predict multiple medical scores (i.e. MMSE, ADAS, and diagnosis feature) by combining multimodal data (i.e. MRI, PET, and CSF). They used SVMs and baseline data from a small number of modalities. Recently, Ding et al. [17] found that most AD studies consider a limited number of features, which are potentially inadequate to understand this complex disease. Yet, despite much ongoing research, predicting AD progression is still a challenge [14]. Deep learning (DL) can enhance the performance of this complex modeling task [34]. However, few studies have explored the role of DL in AD prediction based on time series multimodal multitask data.

1.2. Deep learning based AD prediction

Regular ML has been used in previous literature to predict AD progression. Lu et al. [6] utilized Fluorodeoxyglucose PET (FDG-PET) functional imaging to identify cognitively normal (CN) subjects who will convert to MCI early. The authors built a binary classifier using an incomplete RF–robust SVM approach and achieved 90.53% accuracy. Zhang et al. [3] provided a survey of the predictions from pMCI conversion studies. Existing studies depended on a limited number of biomarkers, which could be insufficient to provide a complete interpretation of the disease. For example, Liu et al. [9] built their model based on MRI and CSF modalities only. Furthermore, most studies currently neglect any temporal dependency within the feature series and across the different features and instead focus on cross-sectional data. Cho, et al. [35] used the ADNI data to predict probable AD conversion using single baseline MRI scans. Moreover, most AD progression studies are classification problems, i.e. categorizing patients into a particular category (e.g. CN, MCI, or AD). Few studies have formulated the problem as a regression task [36–38], also these models fitted logistic or polynomial functions to the longitudinal dynamics of each biomarker separately. These studies have mostly relied on independent biomarker modeling, and no study has considered the temporal dependencies among features [39]. Solving these issues using regular ML classifiers has critical limitations, especially when it comes to learning multiple tasks by fusing many time series modalities together [14]. DL techniques have demonstrated promising prediction results in several areas [40,41]. These techniques are based on recurrent neural networks (RNN) or CNN architectures, which are powerful and can extract deep longitudinal features from fused multivariate time series data [21,27,42]. Tabarestani et al. [43] used two variations of an RNN, namely a long short-term memory (LSTM) and a gated recurrent units (GRU), to predict the patient's status for the next three time points using the previous three historical time points. Alternatively, several studies have formulated AD progression detection as a regression task based on CSs [17,18] considering the fact that CSs are highly predictive factors for AD progression [31,44]. For example, Ding et al. [17] used CDR as an index of AD severity to build a Bayesian network model for AD classification using the AIBL longitudinal dataset. Frisoni et al. [45] found a strong correlation between MRI features and those scores. Most regression studies have focused on predicting one score, based on regular ML algorithms such as SVMs and RF [11,15,22,46–48]. Some studies have set up MCI progression as a multimodal single-task regression model, where one CS as MMSE or ADAS-Cog is used as an indicator for AD progression [10,16]. Recently, Yagi et al. [49] concluded that each CS marker has its limitations, and predicting AD progression based on multiple CSs is medically intuitive and more accurate. Recent works on AD prediction and progression have formulated the problem as a multitask problem, here the model optimizes a multi-objective cost function. Simultaneously learning for multiple related tasks has been proven to perform better than learning for single tasks separately [11,15]. DL techniques are more accurate for jointly learning multiple tasks [34]. In the context of AD, Wang et al. [11] began the trend in AD progression modeling towards DL, multitask, and time series analysis. Liu et al. [30] presented a CNN-based model for concurrently learning the AD diagnosis and CSs regression. Choi et al. [50] also proposed a CNN-based model to detect pMCI patients using their PET images as a single-task model. A multimodal single task classification model was proposed by Spasov et al. [42], where authors utilized the CNN to detect AD progression. In their work, authors combined MRI, demographic, neuropsychological, and APOE e4 genetic data in a late fusion process to do classification. Most Alzheimer's DL models use

CNN models with a single (baseline) MRI scan [18], but these models are less accurate and not medically acceptable [17]. Wang et al. [21] proposed an LSTM-based regression model to predict AD progression based on time series data with non-uniform visit time intervals. Liu et al. [30] proposed a CNN-based model for AD diagnosis and predicting clinical scores. Their model fused MRI data and demographic features of only the baseline visits. Most of the recent Alzheimer's DL models are based on the CNN and consider the baseline MRI scans [18]. Usually medical experts accumulate longitudinal multimodal patient data before making any progression decisions. Therefore, models that only depend on baseline data of the patient are less accurate, less sufficient, and not medically acceptable [17]. Furthermore, most DL models proposed for the AD diagnosis and progression are formulated as binary classification problems, and multiclass fine-grained models are still far from achieving satisfactory results that can be applied clinically [51]. Moreover, DL needs a huge dataset to fit models properly and given the small datasets available in the AD domain, DL is prone to overfitting. DL models are a black-box, and these models that are not interpretable are less acceptable in the medical domain.

1.3. The study hypotheses

The goal of our study is to answer the following research questions: (1) Can the use of fused multimodal time series data and a deep LSTM-based model lead to more accurate prediction of AD progression? (2) Does the joint learning of multiple regression tasks generate a more robust and stable DL model? (3) Does a hybrid model – consisting of a deep learning model for feature learning and an ML model for classification – improve the overall performance of the AD progression detection? (4) As cognitive scores are crucial for AD progression detection, is it possible to predict the future values of a set of cognitive scores and use them to predict AD progression? (5) Does the resulting model become more explainable than the black-box DL models? (6) What is the most important category of MRI features (e.g. volume, cortical thickness, temporal, etc.) in terms of contributing to the most accurate predictions? To the best of our knowledge, there is no current study in the AD domain that critically analyzes these questions. To answer these questions, extensive experiments were carried out based on the AD neuroimaging initiative (ADNI) dataset of 1371 subjects. In this study, we propose a hybrid DL model (HDL) to help overcome the current limitations of DL-based models for AD prediction. The HDL model predicts AD progression after 2.5 years (i.e. at month 48) based on the patient's multimodal time series data from baseline (BL), month 6 (M06), month 12 (M12), and month 18 (M18). The proposed DL model can be used to directly predict AD patients at M48 as a multiclass classification task. This kind of outcome is poorly predicted by simple DL models as shown later in the results section. In the proposed HDL model, a multivariate BiLSTM model is used for deep feature representation learning, and a regular ML model such as decision tree (DT), random forest (RF), or support vector machine (SVM) is used for the classification task. We call this approach deep feature-based learning (DFBL). BiLSTM is a powerful technique for extracting dynamic dependence relationships from longitudinal features, especially from complex and heterogeneous multivariate time series. However, the resulting architecture is complex, and adding Softmax classification to this architecture has certain limitations. In an alternative approach, our DFBL adopted both DL and regular ML models that were trained concurrently. The DL model extracts the deep features from the multimodal data, and these features are used to train the regular machine learning model. This paradigm has been used in other literature to explore the idea of replacing the usual Softmax classifier with other ML classifiers such as SVM or RF. Remarkably, this design has achieved better results than the model it replaced [52–54]. In the context of AD, Zhu et al. [55] extracted representative features from the volume of the gray matter data using a discriminative self-representation sparse regression, these extracted features are fed into an SVM classifier to make the prediction. In [51], Amoroso et al. combined RF for feature extraction and a feed-forward neural network for classification. Lin et al. [56] utilized a CNN along with PCA-LASSO to extract deep features from MRI data, then used SVM to classify the disease. These studies are based on a single modality (i.e. MRI) of a single visit (i.e. no time series), which have several limitations [57]. Unlike the previous literature, we fuse the four time-series modalities from neuropsychological battery markers, CSs sub-scores, MRI, and FDG-PET biomarkers. These modalities are popular in the AD domain [58]. These modalities are fused using a deep BiLSTM model.

We also propose in the study another paradigm called multitask regression-based learning (MRBL) to check the multitask learning capability of the proposed DL models. In the first phase of this design, a multivariate BiLSTM model is used to jointly learn seven regression tasks; then, in the second phase, a regular ML model is used to learn the multiclass AD progression classification based on the predicted CSs and patient demographics. The seven regression tasks are used to predict seven of the most accurate and popular CSs in the AD domain (i.e. ADAS 13, MMES, CDRSB, MOCA, RAVLT, FAQ, ADNI MEM, and RAVLT) [14,29,58,59]. The multitask paradigm works as a regularizer for the seven tasks to efficiently train a DL model than single-task models. CS markers are robust predictors for AD progression [18,44,60]. In other words, based on the temporal changes of CSs, it is possible to predict AD progression quite accurately. In real-world practice, medical experts always rely on CSs to make these decisions [5]. Most previous studies have predicted a single cognitive score using either the multimodal or multitask paradigm based on regular ML techniques [8]. In addition, most of these studies are based on a small number of modalities and use only baseline data [60], they used these predicted values as an indicator for AD progression. Our model jointly predicts seven scores based on a deep LSTM model and multimodal time-series data. This includes multiple longitudinal modalities and simultaneously predicts multiple CSs that might help explain and more accurately predict AD progression [5]. Multitask learning assumes that different tasks have common representation space. Multimodal learning analyzes how multiple sets of clinical data can be merged longitudinally to extract more abstract features on AD progression. In other words, multiple tasks share features from multimodalities to jointly model correlation among outcomes and optimize a multi-objective cost function. In a medical environment, expert users never accept a computerized clinical decision support system unless they understand why and how a certain recommendation is given [61]. In our case, the MRBL improves the interpretability of the DL model. Using an easy to interpret model such as DT in the second phase of MRBL produces an easy to interpret recommendation [62,63]. The resulting explanation is much accurate than using interpretable models to explain the behavior of DL. In the current design, the interpretable model is created based on CSs features learned by the DL model. To extend this idea, we explore two interpretable fuzzy classifiers which are based on Fuzzy Logic (FL), including the fuzzy unordered rule induction algorithm FURIA [29] and a multi-objective evolutionary fuzzy classifier (MOEFC) [61]. To select the best scores, we study the correlation between all CSs and AD progression by using suitable statistical analysis tools and select the highest correlated scores. In addition, the best set of modalities, with the most suitable set of features, was selected and used by the BiLSTM model. In both MRBL and DFBL, each modality is learned by a separate BiLSTM module to extract deep longitudinal features. After extracting longitudinal features from every modality using a separate LSTM model, we use the late fusion technique based on BiLSTM to extract any complementary features from different modalities jointly. From these, the four modules are fused and learned again with another deep BiLSTM module. Predicted makers are integrated with the most progression sensitive baseline features to further predict the exact M48 diagnosis (CN vs. MCI vs. AD). By comparing the classification at M18 with M48, we can determine the progress MCI patients within the 2.5 years from M18. To the best of our knowledge, these explored design paradigms have not been studied in the AD domain, especially for progression detection. As brain atrophy from AD may start as early as twenty years prior to symptoms [57], analysis of MRI data is critical to AD prediction. However, most previous studies have concentrated on exploring the predictive power of specific MRI measures for AD predicting such as hippocampal atrophy, volumetric abnormality, temporal atrophy, cortical thickness abnormality, etc. [13,59,64,65]. As a method of feature selection, we explore the role of different categories of MRI features, in combination with other modalities, on the models' performance.

To sum up, our study makes the following contributions:

- We propose two novel hybrid DL models to tackle the current limitations of DL-based models for AD progression prediction. Our model predicts AD progression 2.5 years later (i.e. at month 48) based on the patient's partial or entire multimodal time series data from the previous 18 months.
- The first HDL model, called DFBL, uses a multivariate BiLSTM architecture for deep feature representation learning, and a regular ML model such as DT, RF, SVM, or FL for the AD prediction task.
- The second HDL model, called MRBL, has two main phases. In the first phase, it uses a multivariate BiLSTM model to jointly learn seven regression tasks; then, in the second phase, a regular ML model is used to learn the multiclass AD progression (CN vs. MCI vs. AD) based on the predicted CSs from the first stage.
- Comprehensive evaluations were conducted based on data from 1371 subjects collected from the ADNI. We explored a set of DL and ML techniques, such as SoftMax classifier, RF, SVM, DT, general model (GM), FURIA, and MOEFC.
- Broad performance comparisons were carried out to show (i) the effect of the included time-series data, when (1) using all of the patient's data (four time-steps), (2) using the BL visit data, and (3) using the M18 visit data, (ii) the effect of applying a single and multitask approach for the MRBL model, (iii) the effect of using BiLSTM architecture against using only flattening time-series data with a conventional DL architecture, and (iv) the role of fusing different MRI feature spaces with other modalities to reduce noise and improve the performance.

The remainder of the paper is organized as follows. Section 2 presents the methods and techniques used in our architecture for deep feature representation and classification. Section 3 presents the details of the proposed hybrid model. The experimental framework that explains the flow of the comprehensive experiments in this paper is provided in Section 4. The results and discussion are presented in Section 5. Finally, Section 6 concludes the paper.

2. Methods

In this section, we introduce the main techniques adopted in our proposed architecture such as LSTM, BiLSTM, and ML algorithms that were optimized to work as classifiers for AD progression detection.

2.1. Long short-term memory

LSTM is a powerful technique to detect patterns in time series and sequential data. However, the number of studies exploiting LSTM for AD multimodal longitudinal data analysis is still limited. AD data from each time point is correlated with the data from the preceding and following time points. To fully capture the temporal variations in correlations of AD data, we apply the BiLSTM, which consists of a forward LSTM ($i = 1, \cdots, n$) and a backward LSTM ($i = n, \cdots, 1$). The BiLSTM networks are used in two positions. First, modalities, i.e. $X \in \{X_{PET}, X_{MRI}, X_{CSD}, X_{NPD}, X_{ASD}\}$ are trained separately by BiLSTM subnetwork. We have five modalities, so we have five two-layer BiLSTM networks trained concurrently and independently to capture the temporal features within single features and among features of different modalities. Second, the learned features from these five networks are fused and learned using another two-layer BiLSTM network to extract the common features among different modalities. Fig. 1 represents the BiLSTM network used to fuse different modalities.

For each time point, the outputs of the forward and backward BiLSTMs are concatenated to form the output of the BiLSTM at that time point, see Fig. 2. We add two BiLSTM layers to efficiently process and summarize longitudinal data of each modality separately. Since the patient time series are only 4 steps, the two layers do not increase the computation loads. The LSTM cell structure consists of three gates: the input gate (i_{t_n}), forget gate (f_{t_n}), and output gate (o_{t_n}). These gates are used to regulate information contained in the cell state. C_{t_n}, $C_{t_{n-1}}$ and $\widehat{C}_{t_n}$ are the cell status contents at the time t_n, the cell status value of the last time step, and the update of the current cell status value, respectively. $h_{t_{n-1}}$ is the value of a memory cell in the hidden layer at the t_{n-1} time step. h_{t_n} is the output value of a hidden layer at the time t_n derived from $\widehat{C}_{t_n}$ and $C_{t_{n-1}}$. The calculated weight matrices and biases vectors are represented as θs and b, respectively. These are updated following the back propagation algorithm. $\otimes$ is the Hadamard product; $\otimes$ is the concatenation operator; σ is the logistic sigmoid function; ϕ is the activation function output, e.g. *SoftMax*, *ReLU*, or *Tanh*. The Eqs. (1) to (7) represent the flow of information in an LSTM memory cell at each step.

$$f_{t_n} = \sigma\left(\theta_f \bullet \left[h_{t_{n-1}}, x_{t_n}\right] + b_f\right) \tag{1}$$

$$i_{t_n} = \sigma\left(\theta_i \bullet \left[h_{t_{n-1}}, x_{t_n}\right] + b_i\right) \tag{2}$$

$$\tilde{C}_{t_n} = tanh\left(\theta_C \bullet \left[h_{t_{n-1}}, x_{t_n}\right] + b_C\right) \tag{3}$$

$$C_{t_n} = \left(f_{t_n} \otimes\ C_{t_{n-1}} \oplus i_{t_n} \otimes\ \tilde{C}_{t_n}\right) \tag{4}$$

$$o_{t_n} = \sigma\left(\theta_o \bullet \left[h_{t_{n-1}}, x_{t_n}\right] + b_o\right) \tag{5}$$

$$h_{t_n} = o_{t_n} \otimes \tanh\left(C_{t_n}\right) \tag{6}$$

$$y_n = \varphi(\theta_y h_{t_n} + b_y) \tag{7}$$

The LSTM unit captures information related to the previous input sequence in a time series but does not capture the relation between future and previous sequences. BiLSTM [66] merge two independent hidden LSTM layers in the forward and backword directions into the same output to learn the overall dependencies in a time series. BiLSTM gets an input sequence, $X = (X_{t0}, X_{t1}, \ \ldots, X_{tn})$, in a forward hidden sequence as in the conventional LSTM, $\overrightarrow{h}_t = (\overrightarrow{h}_{t0}, \overrightarrow{h}_{t1}, \ldots, \overrightarrow{h}_{tn})$, and a backward hidden sequence in the opposite direction, $\overleftarrow{h}_t = (\overleftarrow{h}_{t0}, \overleftarrow{h}_{t1}, \ldots, \overleftarrow{h}_{tn})$. The output vector of a hidden layer $y_t =$

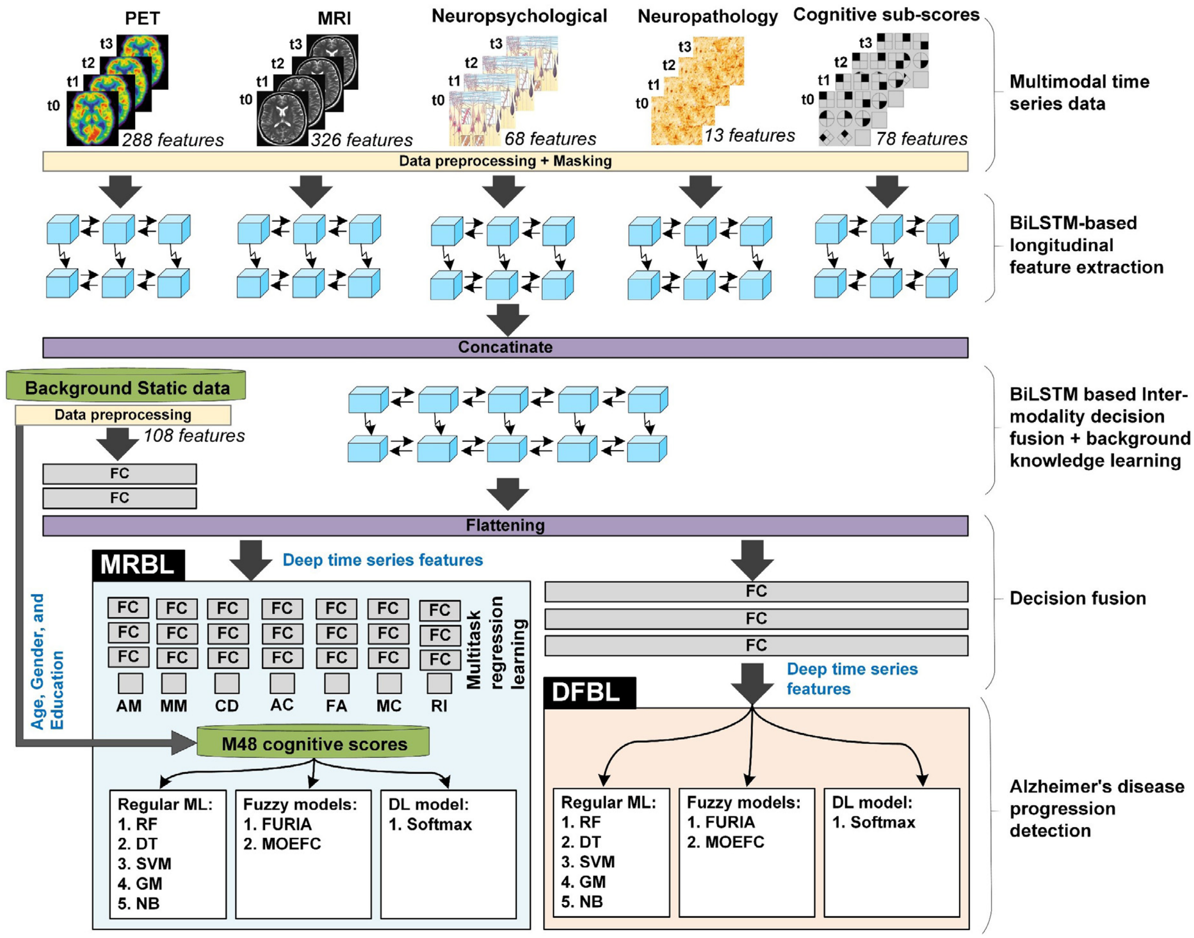


Fig. 1. BiLSTM-based framework for time series data learning and multitask progression prediction.

$(y_{t0}, y_{t1}, \ldots, y_{tn})$, $t = 1, 2..t$ is the combination of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$, $y_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$, as shown in Eqs. (8) to (11).

$$\overrightarrow{h}_{tn} = \sigma\left(\theta_{\overrightarrow{h}_n} \bullet \left[\overrightarrow{h}_{tn-1}, x_{t_n}\right] + b_{\overrightarrow{h}_n}\right) \tag{8}$$

$$\overleftarrow{h}_{tn} = \sigma\left(\theta_{\overleftarrow{h}_n} \bullet \left[\overleftarrow{h}_{tn-1}, x_{t_n}\right] + b_{\overleftarrow{h}_n}\right) \tag{9}$$

$$\left(\overrightarrow{h}_{t0}, \overleftarrow{h}_{t0}\right) \ldots \left(\overrightarrow{h}_{tn}, \overleftarrow{h}_{tn}\right) = BiLSTM(X_{t0}, X_{t1}, \ \ldots, X_{tn}) \tag{10}$$

$$y_t = \sigma\left(\theta_{y_t \overrightarrow{h}_n} \overrightarrow{h}_{tn} + \theta_{y_t \overleftarrow{h}_n} \overleftarrow{h}_{tn} + b_{y_t}\right) \tag{11}$$

The output of a BiLSTM y_t is fed to the next hidden layer. The output y_t from the former layer is fed as an input to the later layer, see Fig. 1. Since the time series of our study are not long, the computation load of the two BiLSTM layers is insignificant.

2.2. Regular machine learning models

Many regular ML algorithms have been optimized to work as classifiers based on either learned deep features from the LSTM model or the predicted cognitive scores of M48. These models include SVM, RF, DT, Naïve Bayes (NB), and GM. According to our results, we found that the best model is RF [67]. In addition, fuzzy classifiers are popular classification approaches especially in the medical domain, because (1) they provide interpretable models based on linguistic labels in their fuzzy rules, (2) they are able to handle the vague and uncertain nature of medical data, and (3) they can be trained easier than deep learning models [68,69]. More details about the regular ML models used can be found in Supplementary File 2.

3. Proposed learning approach

In this section, we describe the detailed steps of the proposed hybrid multitask LSTM-based models. The model predicts AD progression at M48 based on four time-steps (BL, M06, M12, and M18). As shown in Fig. 1, the inputs to the system are two types of data, namely four time-series modalities and baseline data. In the first layer, each time series modality is learned by a deep LSTM module. In the second layer, the learned deep features of the first layer are fused and used by another multilayer LSTM module to learn deeper features. The baseline data includes many critical features collected from the first visit including age, education, TAU, APOE, etc. These data work as background knowledge in the DL model. These data are learned by using a feed forward neural network (FFNN). Learned features from the FFNN and second layer of LSTM are fused again and used in the two different ML models. We compare two design schemes: DFBL and MRBL. In the DFBL design, these fused features are concurrently used to train regular ML models (RF, DT, SVM, GM, and NB), FL models (FURIA and MOEFC), and a DL model (SoftMax). In the MRBL design, these fused features are used to train a multitask model. We use the learned deep features to learn seven regression tasks based on a set of dense layers for each task, as shown in Fig. 1. Each regression task is responsible for predicting one CSs at M48. We create a new dataset of the predicted CSs by adding another three critical features (i.e. age, gender, and education). The resulting dataset is used to train another multiclass classifier (CN vs. MCI vs. AD) to predict AD progression at M48. We tested many classifiers for this step including regular ML models (RF,

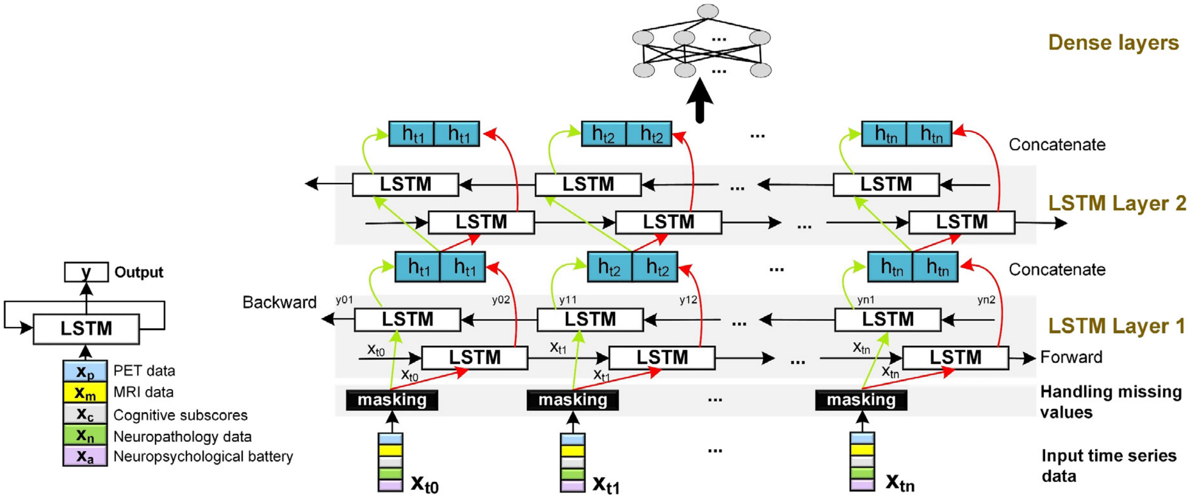

Dense layers
Concatenate
LSTM Layer 2
Concatenate
LSTM Layer 1
Backward
Forward
Handling missing values
Input time series data
masking
LSTM
h_{t1} h_{t2} h_{tn}
y_{01} y_{02} y_{11} y_{12} y_{n1} y_{n2}
x_{t0} x_{t1} x_{tn}
y Output
x_p PET data
x_m MRI data
x_c Cognitive subscores
x_n Neuropathology data
x_a Neuropsychological battery

Fig. 2. The architecture of the proposed LSTM layer for AD prediction.

DT, SVM, GM, and NB), FL models (FURIA and MOEFC), and DL model (SoftMax). The multitask learning process is formulated as discussed in the next section.

3.1. Multitask LSTM-based modeling

The proposed DL framework is a multimodal multitask model, where the model simultaneously learns relevant tasks Y based on AD modalities M. The modalities set M can be depicted as X={$X^{(1)}, \ldots, X^{(M)}$}, and the tasks to be leaned represented as $Y = \{y^{(1)}, \ldots, y^{(T)}\}$, where $y^{(j)} = \{y_1^{(j)}, \ldots, y_N^{(j)}\}$ is a vector of N values for N subjects. A modality $X^m \in X$ is represented as $X^m = \{x_1^{(m)}, \ldots, x_i^{(m)}, \ldots, x_N^{(m)}\}$ of N patient examples, where an example $x_i^{(m)}$ is a multivariate time series $x_i^{(m)} = \{x_{i1_t}^{(m)}, x_{i2_t}^{(m)}, \ldots, x_{if_t}^{(m)}\}$, for $t = 1, \ldots, s$ time-steps and f set of univariate time series. Each patient $i \in N$ is represented as $x_i = \{x_i^{(1)}, \ldots, x_i^{(m)}, \ldots, x_i^{(M)}, y_i^1, y_i^2, \ldots, y_i^T\}$, where $i = 1, \ldots, N$, and y_i^1 is the label of the first task for the ith example. For simplicity, Fig. 3 illustrates a single modality for N subjects. Each modality has four-time steps and M features, that is, input X, and four CSs as regression tasks Y. The model should optimize shared parameters (θ^{sh}) and task-specific (θ^t) parameters, where the parametric hypothesis of each task is $f^t\left(x, \theta^{sh}, \theta^t\right) : X \to \mathrm{y}^{(t)}$, and the task-specific loss functions are $\mathcal{L}^t\,(.,.) : \mathrm{y}^{(t)} \times \mathrm{y}^{(t)} \to \mathbb{R}^+$. In our proposed multimodal multitask model, the learning process is a gradient-based multi-objective optimization of task-specific losses, as shown in Eq. (12).

$$\begin{aligned}\min_{\substack{\theta^{sh},\\ \theta^1,\ldots,\theta^T}} \; & L\left(\theta^{sh}, \theta^1, \ldots, \theta^T\right) \\ = \min_{\substack{\theta^{sh},\\ \theta^1,\ldots,\theta^T}} \; & \left(\widehat{\mathcal{L}}^1\left(\theta^{sh}, \theta^1\right), \ldots, \widehat{\mathcal{L}}^T\left(\theta^{sh}, \theta^T\right)\right)\end{aligned} \tag{12}$$

where $\widehat{\mathcal{L}}^t(.,.)$ is a task-specific loss function defined as $\widehat{\mathcal{L}}^t\left(\theta^{sh},\ \theta^t\right) = \frac{1}{N}\sum_i \mathcal{L}(f^t\left(x_i; \theta^{sh}, \theta^t\right), \mathrm{y}_i^{(t)})$. Our model predicts seven regression tasks. The regression tasks are equally treated with the objective function defined in Eq. (13), for m, which is the number of θ^{sh} and θ^t parameters.

$$\widehat{\mathcal{L}}^t\left(\theta^{sh},\ \theta^t\right) = \frac{1}{T}\sum_{t=2}^{T}\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{y}_i^{(t)} - \widehat{\mathrm{y}}_i^{(t)}\right)^2 + \frac{\lambda}{m}\sum_{j=1}^{m}\theta_j^2 \tag{13}$$

where the loss function is the mean squared loss for regression, and the added term in the equation is the regularization term. In this study, the T cognitive scores are exploited in the back propagation algorithm to update weighs of the BiLSTM layers, and learn afterword the most relevant features in the dense layers. It is worth noting that the DL model can simultaneously optimize multiple tasks if they share inputs [70].

3.2. Data preparation

Prior to the training of the proposed model, several data preparation steps were carried out. The first step is to determine the best CSs that can be used to predict AD progression at M48. Based on previous literature and our statistical significance analysis of most CSs (Supplementary File 2), we determined the discriminative power of each independent score for measuring AD severity. We use the correlation analysis, non-parametric Kruskal–Wallis test, and probability density function to achieve this task. Selected scores are used as the regression features in the proposed model. Regarding the multimodal time series data and the baseline data, a set of preprocessing steps is performed to improve the data quality. These steps include the following:

- In the first step, missing data are handled for both time series modalities and baseline static data. For static data, all features with more than 30% missing data are removed. For the rest of the features with missing data , we use the k-nearest neighbors (KNN) algorithm to impute a missing value with the average of its k neighbors, assuming that the missing values are at random places. In our study, the mixed Euclidean distance function is used, k is set to 10 empirically via experiment. The average percentage of missing values handled by KNN imputation is 9.6% (8.64% in training data and 0.96% in testing data). Although the baseline static data (e.g. Tau, Amyloid Beta, age, etc.) are: (1) medically critical for the medical experts to infer about the patient's status, and (2) valuable to enhance the performance and robustness of the deep learning models as illustrated in Figure S4 of Supplementary File 2, we found that the use of different imputation techniques has a minor effect on the overall model performance. For time-series data, any univariate time series time with either more than 30% missing data or has no baseline reading is removed. Handling missing values in time series is conducted in two sequential strategies. (1) Unify all time series by filling values at a specific time based on our knowledge of ADNI procedures. Some time series values are intentionally not collated by ADNI based on the patient status. For example, CN patients of ADNI1 do not have an MRI scan at the M18 visit. Also several lab

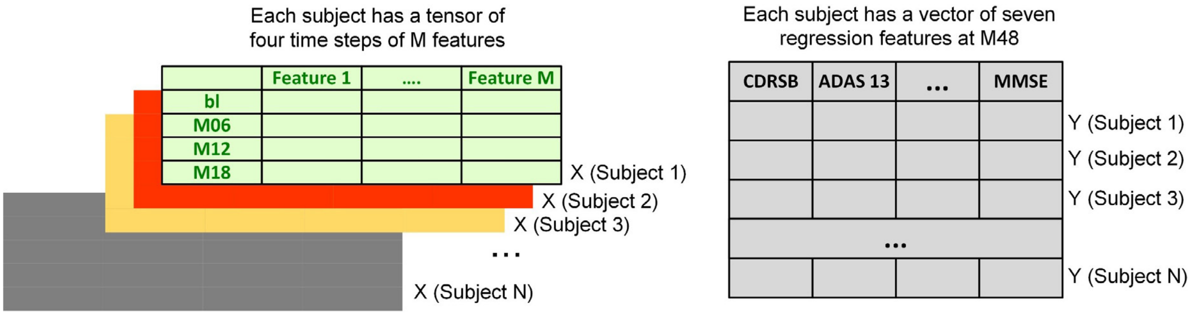


Fig. 3. The structure of each time series modality for the LSTM model.

and cognitive tests as well as neuroimaging scans are not done periodically for all patients. We follow a restricted procedure to fill these values by a forward filling of former values when the diagnosis has not changed. We only consider a value is missing in the time series if the patient diagnosis has changed. This technique of unifying the time series is common in Alzheimer's literature [71]. (2) We take the advantages of LSTM's ability to analyze time series with variable lengths or has missing data at certain time points. An LSTM masking layer is added after the input to handle the missing inputs. This method is well known in the literature and outperform other imputations techniques for handling missing values of time series data [72–74]. In our model, the missing time points are filled with meaningless values (−55) in the masking layer. The resulting time-series data have a regular interval of six months and fed to our BiLSTM model directly.

- The second step is to divide the dataset into a training set (90%) and a test set (10%). The training set is for building and validating the model, while the test set is for checking the generalizability of the model. This process is repeated ten times and the average performance is reported.
- The third step is to rescale the data to be homogeneous. We normalize the data in the range [0, 1]. Outliers are detected and replaced by the average value for its class.
- The fourth step is data balancing. Our dataset is not severely imbalanced because the classes distribution are: CN (30.56%), MCI (44.71%), and AD (24.73%). Since the ML models are prone to be biased in the case of imbalanced datasets. There are many techniques for handling imbalanced datasets, and the synthetic minority oversampling technique (SMOTE) is a popular technique for oversampling and undersampling data [75]. In our case, SMOTE is adopted to address the class imbalance in the training set by only using the oversampling technique.
- Feature selection is a critical step in the ML pipeline. In the embedding methodology, selection is part of the learning procedure. We are mainly rely on the power of DL to extract the most informative features from the time series data. In addition, we explicitly examine the role of different MRI feature categories including hippocampal, volume, temporal, cortical thickness, etc.

After data preparation, the DL model is trained and validated. To guarantee unbiased tuning of model hyperparameters and because our dataset is relatively small, the LSTM model training and validation process (i.e. hyperparameters optimization) are based on stratified 10 fold cross-validation repeated 10 times [76]. The DL model is separately optimized for both design schemes (MRBL and DFBL). The tuned models are tested using the unseen test sets. Keeping the test set untouched is critical to estimate the generalization performance of the selected model. It is worth noting that the final evaluation of the resulting model is performed on the held-out set, which is not used prior to either model training or tuning of its parameters. This process is repeated 10 times, and the average performance is reported. For the DFBL design, the hybrid model is optimized concurrently. The LSTM model and the connected classifier (ML, FL, SoftMax) are trained simultaneously. For the MRBL design, the LSTM model is optimized for learning the seven regression tasks, and the predicted CSs are collected. A separate process for optimizing the classifier (ML, FL, SoftMax) is carried out based on the resulting feature set.

4. Experimental framework

4.1. ADNI study

The data used to prepare this study was obtained from the Alzheimer's disease neuroimaging initiative (ADNI) database. The data was accessed on March 18, 2019. We selected 1371 subjects (54.5% male) from ADNI 1, ADNI GO, and ADNI 2 based on the availability of data. Based on the subjects diagnosis from the baseline to month 48 of ADNI study, patients are grouped into 4 classes, as shown in Fig. 4. (1) 419 subjects are diagnosed as CN and remained CN during the study period (i.e., from the baseline to M48). (2) 473 subjects are diagnosed as MCI during all the study period. (3) 140 subjects are diagnosed as MCI at baseline+M06+M12+M18 visits and progressed to AD within 2.5 years starting from M18 (i.e., from M18 up to M48). (4) 339 subjects are diagnosed as AD in all visits. The used list of patient IDs can be found in Supplementary File 2. As shown in Fig. 4, no subjects with reverse diagnosis, (e.g., from MCI to CN, or AD to MCI) during visits, are included from the study. In other words, we considered the subjects with the reverse diagnosis as misdiagnosed, this comes from the consideration that AD is an irreversible form of dementia. Furthermore, all subject that have converted from CN to AD are also excluded. Our collected data has two main categories. Several neuropsychological tests and biomarkers have been collected and individually validated for AD progression. In this study, we will fuse two types of data. The first type is time-series data including the five modalities of CSs subscores (78 features), neuropsychological battery (55 features), neuropathology data (13 features), MRI (326 features), and PET data (46 features). Specific categories of MRI that are explored include volumes (115 features), cortical thicknesses (144 features), hippocampus region (17 features), surface areas (75 features), temporal region (47 features). The second type is baseline (i.e. not time series) data including the 108 features of CSF markers (i.e. amyloid-β peptide of 42 amino acids-Aβ1–42 [ABETA], TAU, and phosphorylated TAU [PTAU]); genetic information APOE4; demographics; medical history; symptoms; lab tests; and physical examinations. A full description of our dataset can be found in Supplementary file 2. Table 1 shows demographic and clinical information of the subjects.

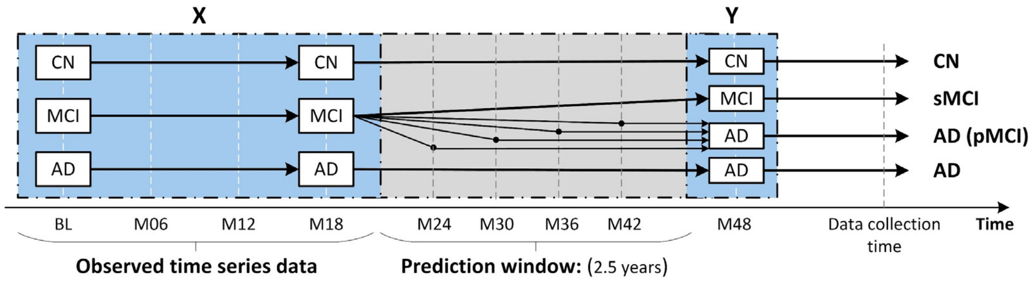


Fig. 4. Data formulation process.

Table 1
Descriptive statistics of the main features of the dataset at baseline and month 84.

	CN		MCI (n = 613)				AD (n = 339)	
	(n = 419)		sMCI (n = 473)		pMCI (n = 140)			
	Baseline	M84	Baseline	M84	Baseline	M84	Baseline	M84
Gender (M/F)	191/228	191/228	283/190	283/190	74/66	74/66	187/152	187/152
Age (years)	73.97 ± 05.74	73.97 ± 05.74	72.92 ± 07.75	72.92 ± 07.75	74.34 ± 07.14	74.34 ± 07.14	75.00 ± 07.80	75.00 ± 07.80
Education (y)	16.39 ± 02.69	16.39 ± 02.69	15.79 ± 02.96	15.79 ± 02.96	16.14 ± 02.94	16.14 ± 02.94	15.12 ± 02.98	15.12 ± 02.98
FAQ	00.19 ± 00.73	00.30 ± 01.11	02.09 ± 03.14	03.47 ± 04.81	04.55 ± 04.53	15.04 ± 07.52	13.30 ± 06.85	18.69 ± 07.37
MMSE	28.98 ± 01.16	28.92 ± 01.31	27.63 ± 02.13	27.06 ± 02.72	26.45 ± 02.09	22.23 ± 04.01	22.06 ± 03.79	19.46 ± 05.44
MoCA	25.67 ± 01.96	25.30 ± 02.54	23.14 ± 02.69	22.74 ± 03.14	21.77 ± 01.99	17.83 ± 04.47	17.48 ± 03.54	16.07 ± 05.30
APOe4	00.27 ± 00.48	00.27 ± 00.48	00.51 ± 00.66	00.51 ± 00.66	00.84 ± 00.69	00.84 ± 00.69	00.85 ± 00.71	00.85 ± 00.71
ADAS 11	05.63 ± 02.82	05.75 ± 03.24	08.89 ± 3.80	10.55 ± 06.44	13.41 ± 05.02	22.95 ± 11.58	20.21 ± 06.53	25.84 ± 12.12
ADAS 13	08.70 ± 04.09	08.96 ± 04.45	14.78 ± 05.86	16.22 ± 08.77	19.61 ± 05.04	30.43 ± 10.17	29.99 ± 07.99	37.00 ± 13.08
RAVLT imm.	45.79 ± 09.71	45.78 ± 10.85	36.42 ± 10.66	34.10 ± 11.81	29.69 ± 07.08	21.72 ± 07.54	22.64 ± 07.48	18.33 ± 08.44
RAVLT learn	06.12 ± 02.23	05.76 ± 02.40	04.56 ± 02.51	03.20 ± 02.54	02.67 ± 02.48	01.48 ± 01.36	01.37 ± 01.35	01.40 ± 01.67
RAVLT forget	03.56 ± 02.83	03.39 ± 02.90	04.42 ± 02.61	04.52 ± 02.39	05.60 ± 02.35	04.72 ± 02.14	04.80 ± 02.83	04.98 ± 02.18
RAVLT % forget	32.45 ± 25.79	33.97 ± 27.60	54.65 ± 30.45	63.87 ± 32.11	76.87 ± 27.30	89.85 ± 32.55	89.50 ± 20.86	92.41 ± 24.20
CDR	00.08 ± 0.31	00.11 ± 00.37	01.34 ± 00.89	01.66 ± 01.29	02.70 ± 01.00	05.35 ± 02.86	05.22 ± 02.28	07.17 ± 03.56

Statistical measures of the features are shown as mean ± standard deviation.

Table 2
Optimized hyperparameters for ML algorithms used in the DFBL and MRBL models.

Model.	Algorithm	Hyperparameter
DFBL	SVM	Support Vectors = 74, Cost = 10, solver – lbfgs, gamma = 0.01, learning_rate = adaptive
	NB	Laplace smoothing = active
	RF	criterion = gain ratio, max_depth = 2, number of trees = 100
	GM	alpha = 0.5, lambda = 9.145E−4, kernel = guession
	DT	criterion = gini, max_depth = 4, minimum leaf split = 2
	FURIA	fuzzy operator = product T-norm, batch = 20, folds = 4, optimization = 2, minimum weight = 2
	MOEFC	algorithm = NSGA2, batch = 20, generations = 100, population = 150, max similarity = 0.4
MRBL	SVM	Support Vectors = 465, Cost = 1000, solver – lbfgs, gamma = 1.0E−4, learning_rate = adaptive
	NB	Laplace smoothing = active
	RF	criterion = gain ratio, max_depth = 4, number of trees = 60
	GM	alpha = 0.5, lambda = 6.855E−4, kernel = guession
	DT	criterion = gain ratio, max_depth = 4, minimum leaf split = 2
	FURIA	fuzzy operator=product T-norm, batch = 20, folds=4, optimization = 2, minimum weight = 2
	MOEFC	algorithm=NSGA2, batch = 20, generations=100, population = 150, max similarity = 0.4

4.2. Model training process

All the experiments of the paper are conducted on a machine with an Intel® Xeon(R) CPU E5-2620 v3 @ 2.40 GHz× 24 with Cuda-10.0 and three GEFORCE GTX TITANx 12 GB GPUs; we used Python 3.7.7 distributed in Anaconda 4.8.3 (64-bit). All the proposed models are implemented using Keras library based on TensorFlow as the backend. For the classification task, the SoftMax activation function with categorical cross-entropy loss was used, while the sigmoid activation function with mean square error loss was applied for the regression tasks. For the learning rate optimization, all deep learning models are adopted Adam optimizer with a learning rate of 0.0001 [77]. The training batch size and number of epochs were 15 and 60, respectively. To speed up, we parallelized the training process across the GPUs. The training of the proposed models is an optimization process to find the best parameters that give the best performance either in the classification task or in multitask regression tasks. We split our dataset into stratified 90% training and validation, and 10% test datasets. We used a procedure known as stratification to randomize the instances at each execution to ensure that training and testing datasets contain a similar proportion of the all classes. This procedure is repeated ten times in all reported experiments to avoid bias. For the DFBL design, the DL model optimizes a multiclass classification task using SoftMax at the output layer. After the optimization process, the model is frozen and used to identify deep features in the test dataset. The resulting deep features are fed to either SoftMax, regular ML, or fuzzy classifiers. For the MRBL model, the DL model is optimized to learn seven regression tasks using the Sigmoid activation function at the output layers. The hyperparameters of the final models are evaluated using stratified 10-fold cross-validation before we decided on the final hyperparameters. In our experiments, we feed each modality of our five modality data to a masking layer followed by two stacked BiLSTM layers. These layers have (128 × 2) and (64 × 2) units with L2 regularization of 0.005, a dropout of 0.15, and a Tanh activation function. Each BiLSTM module has an equal number of LSTM layers, but their weights are independently optimized. The five BiLSTM blocks are fused to a new BiLSTM layer with (320 × 2) units and this is merged with the background static data to learn the more abstract deep features of all modalities. The background data are added to the model after passing two sequential feed-forward dense layers, these reduce the dimensionality of features from 108 to 64. The learned abstract features fed to either the regression or classification task after passing through three consecutive dense layers (64, 32, 32 units for each layer) using the ReLU activation function, L2 regularization of 0.005, and drop out of 0.20. All outputs from the last dense layer are used either to optimize DFBL or MRBL, as illustrated in Fig. 1. For the DFBL architecture, the 32 deep features generated by the last dense layer are fed to three classifies types, including the SoftMax, ML (SVM, RF, DT, NB, and GM), and FL (FURIA, and MOEFC) classifiers to predict the AD progression at M48, see Supplementary Figure S1. The 10-fold cross-validation was used in all the classification experiments with the grid search method chosen to find the optimal hyperparameters. The optimized ML model was evaluated on an unseen test set, the average of the ten evaluations is reported. For the MRBL design, the model jointly optimizes seven regression tasks. The concatenated features of the last BiLSTM layer with a static layer are fed to seven separate blocks, where each block is responsible for one regression task, as illustrated in Fig. 1. The added dense block has three dense layers using the ReLU activation function, L2 regularization of 0.005, and drop out of 0.20. Another dense layer is added for each regression task using a sigmoid activation function, L2 regularization of 0.005, and drop out of 0.20 to predict the regression values, see Supplementary Figure S2. The regression tasks are stopped early when the error does not decrease within the next 30 epochs. The network is trained only on a training set, while validation samples are used to determine when to stop the optimization process. The average results of the regression multitask are reported for the unseen test sets. Similar to DFBL, all the predicted values of the multitask regression are fed to three classifies types, the SoftMax, ML (SVM, RF, DT, NB, and GM), and FL (FURIA, and MOEFC) classifiers, to predict the AD progression at M48. In the MRBL experiments, the 10-fold cross-validation was used in all the classification experiments using the grid search method to find the optimal hyperparameters. The optimized ML model is evaluated on the unseen test set, and the average of ten evaluations is reported. Table 2 illustrates the optimal hyperparameters of the ML algorithms for the DFBL and MRBL models.

4.3. Performance evaluation metrics

The performance in regression tasks is evaluated using the mean absolute error (MAE), as defined in Eq. (14), where N is the number of cases, y_i is the actual value for example i, and $f(x_i)$ is its predicted value.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i)) \tag{14}$$

The performance in classification tasks is evaluated using the standard four metrics of accuracy, precision, recall, and F1-score (Eqs. (15) to (18)), where TP stand for true positive, TN means true negative, FP is false positive, and FN is false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

$$Recall = \frac{TP}{TN + FN} \tag{17}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

4.4. Statistical analysis

4.4.1. Selecting efficient cognitive markers

To determine the independent CSs with the highest discriminative power, we performed the following steps. First, we surveyed the literature to collect the most important cognitive markers. Next, based on our training dataset, we performed a deep statistical analysis to select the most significant scores for use in regression tasks with the DL model. A P-value < .05 is considered significant. We use a set of statistical tests based on the nature of the data, i.e. normal distribution and outliers. The sizes of the different diagnostic groups (i.e. CN, MCI, and AD) are large enough for all statistical tests, and CSs, to have no outliers. For CSs displaying normal distribution including ADAS-cog, MoCA, ADNI MEM, and RAVLT-Immediate, we used a one-way analysis of variance (ANOVA) parametric test to check that the three independent classes are significantly different (i.e. CN-MCI-AD). For post-hoc testing after AVOVA, we use the conservative Scheffe test to check each pair of diagnostic groups (i.e. CN-MCI, CN-AD, and MCI-AD), this prevents type 1 errors. For the groups that had a non-normal distribution including MMSE, FAQ, and CDR we use the non-parametric Kruskal–Wallis test for the three groups tests followed by Dunn's test for post-hoc multiple comparisons based on Bonferroni's correction. Seven markers were selected to be predicted by the LSTM model. As asserted in Fig. 5, statistical analysis indicates that the selected scores are statistically highly significant (P<.0001), and there exists statistically high significant differences among CN and MCI, CN and AD, and MCI and AD (P<.0001). As can be seen in the distributions in Fig. 5, there are significant differences among the means of different groups for each score. As a result, the selected markers are considered as significant predictors of AD progression. Longitudinal changes in the selected CSs could track AD patient progression. As shown in Fig. 6, different Alzheimer's disease classes (i.e. CN, sMCI, pMCI, and AD) have significantly different ranges of values (P<.005) at each specific visit at BL, M06, M012, M18, and M48. As a result, the LSTM model is able to differentiate each class at each visit and track its change from the previous visit. In addition, the pMCI class has a clear progression behavior between M18 and M48 for all selected CSs. In other words, in the M18 visits, all pMCI cases are considered MCI, but in M48 visits these cases have progressed to AD class. We notice that different classes have different levels of progression for each marker. As shown by the standard deviations, the CN and sMCI classes are approximately stable because they are changing very slowly. On the other hand, the pMCI and AD classes are changing fast. For example, CDR has average values of 0.083±0.30, 1.34±0.89, 2.07±1.00, and 5.23±2.29 for respective CN, sMCI, pMCI, AD classes at the BL visit; 0.083±0.30, 1.34±0.89, 2.07±1.00, and 5.23±2.29 at M06; 0.099±0.32, 1.39±0.98, 2.29±1.11, and 6.00±2.93 at M12; 0.099±0.32, 1.38±1.00, 2.48±1.15, and 6.02±2.97 at M18; 0.11±0.37, 1.66±1.30, 5.37±2.69, and 7.18±3.57 at M48. CN and sMCI classes exhibit a slow rate of cognitive decline; however, the pMCI and AD groups undergo statistically significantly increases (P<.001) especially between M18 and M48, indicating a cognitive decline of the participants.

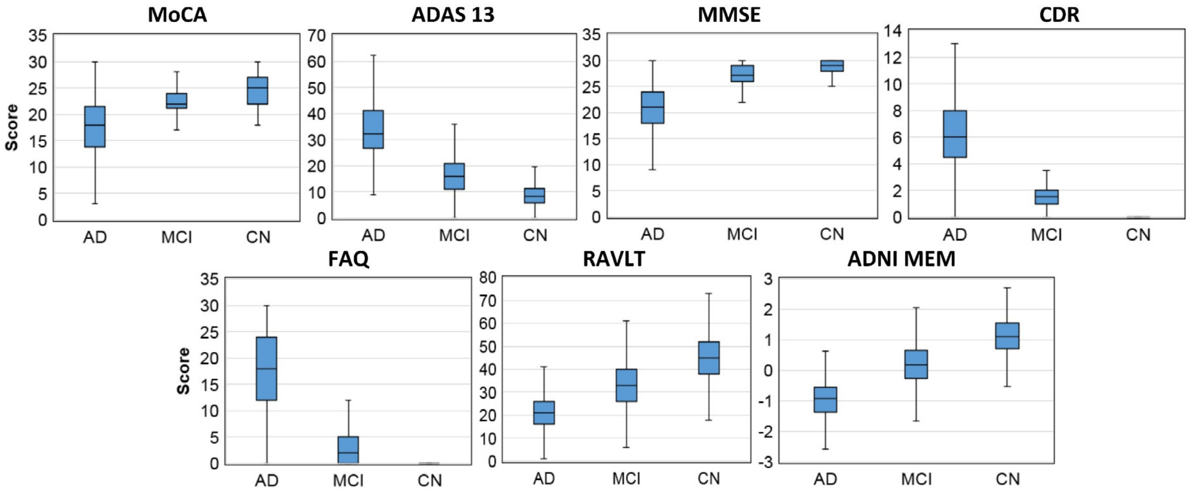


Fig. 5. Group of box plots showing the distributions of seven CSs selected to be learned by the DL model.

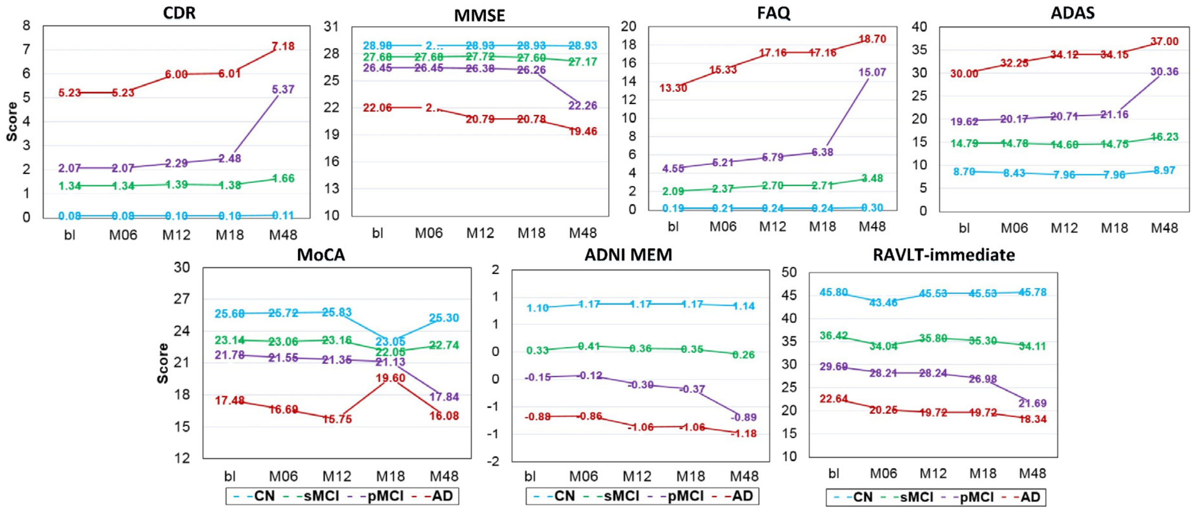


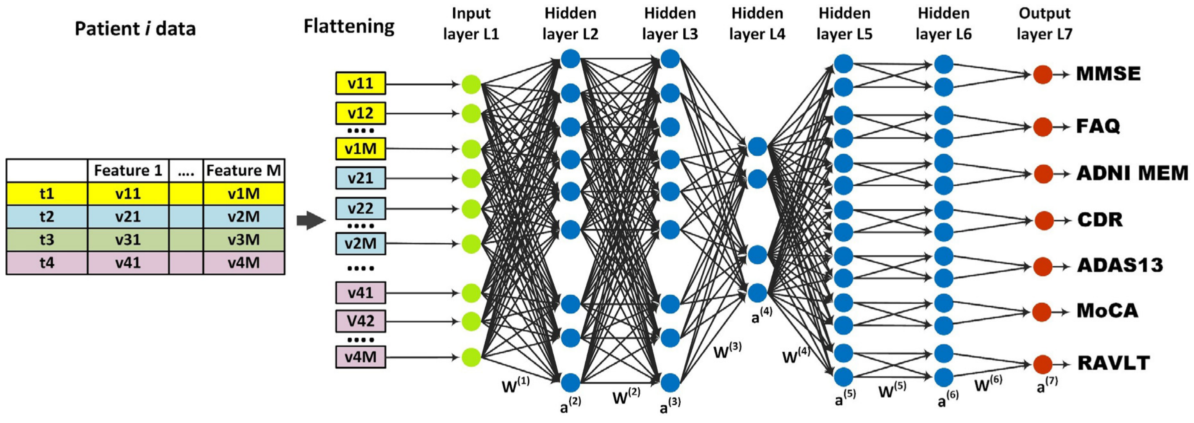
Fig. 6. Patterns of decline of different markers to show the temporal progression of selected CSs from BL reading to M48 reading.

4.4.2. Selecting static features

The selected CS markers in the previous step are used in the LSTM model as the seven regression tasks. The LSTM function is to predict the values of these scores at M48 visit based on the four-time steps of BL, M06, M12, and M18. Other features like age (P<.0005, Student's T-test), gender (P< .0002, Chi-Square test), and the number of education years (P<.0001, Student's T-test) are found to make a significant difference between different classes. These three features are combined with the predicted values of the seven CS markers. These 10 features are used later by the interpretable models to elucidate the model's decisions at M48. The LSTM model combines a large set of static baseline features with five complementary time-series modalities including CSs sub-scores, a neuropsychological battery, neuropathology, MRI, and PET to optimize the multitask (i.e. seven task) regression problem.

5. Results and discussion

To evaluate the performance and effectiveness of our proposed multimodal multitask DL methods, we tested and compared many schemes, as illustrated in Figure S3 of Supplementary File 2. All reported results are based on the unseen testing data. All experiments were repeated ten times and the average performance is reported. In these experiments, we answer the proposed hypotheses in the Introduction Section. Before starting the experiments to evaluate the proposed models, we have conducted an initial evaluation for selecting the best modalities that can be used in our multimodal models, as illustrated in Figure S4 and S5 of Supplementary File 2. We have noticed that including all modalities achieved the best performance and made the model robust and stable. Moreover, we explored the option of replacing BiLSTM layers of the two proposed model of Fig. 2 by BiGRU layers and found ≈ 3%, 2% accuracy degradation in the MRBL and DFBL models, respectively (see Figure S4 and S5 of Supplementary File 2). In Experiment 1, we check the first argument o "*can the use of fused multimodal time series data and deep BiLSTM model lead to a more accurate prediction of AD progression?*" In this experiment, we compare the deep BiLSTM model with the FFNN. The FFNN has been optimized using the BL, M18, and flattened data. The flattening process is inspired by the image processing field. As shown in Fig. 7, for each patient, all time-series data are converted into a single vector. This vector is fed as input to the FFNN. The question of "*does a hybrid model of a deep learning model for feature representation learning and an accurate ML model for classification improve the performance of the resulting DL model?*" is answered in Experiment 2. The question of "*does the joint learning of multiple regression tasks generate a more robust and stable DL model?*" is answered in Experiment 3. The question "*is it possible*



	Feature 1	....	Feature M
t1	v11		v1M
t2	v21		v2M
t3	v31		v3M
t4	v41		v4M

Fig. 7. Time series flattening process and learning with FFNN.

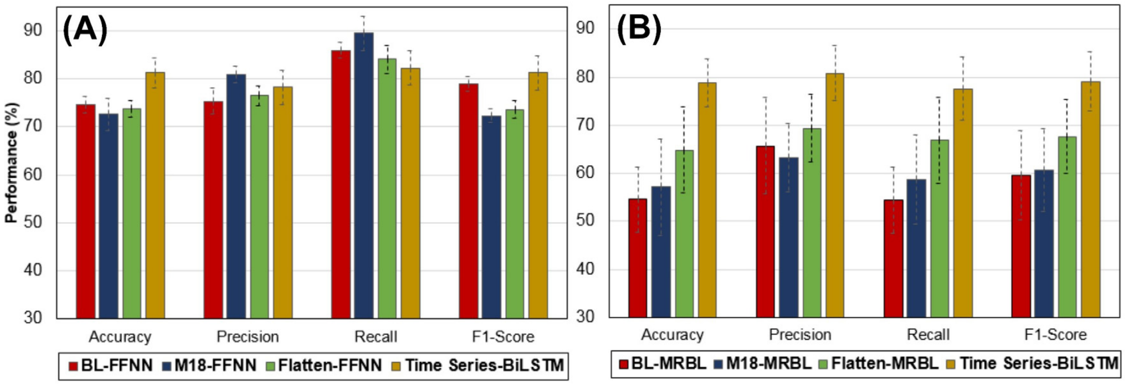


Fig. 8. Comparison between performance for different deep features learning methods.

to predict the future values of a set of cognitive scores and use them to predict the AD progression?" is investigated in Experiment 4. The explainability question is tackled in Experiment 5. Finally, the selection of the best category of MRI features is investigated in every experiment to determine the most critical feature set from the volume, cortical thickness, hippocampus, surface area, temporal region, or all MRI features. Please note that we used this categorization according to previous studies [13,59,64,65]. We select the best category results using our less complex, faster, and more interpretable model.

5.1. Experiment 1: Deep feature extraction

This experiment explores the role of time-series data and the BiLSTM DL model to recognize deep representation in the multivariate longitudinal data. We compare the LSTM model based on multimodal time series data with the FFNN based on the BL, M18, or flattened time series data. Please note that we used the BL data because most previous studies were based on baseline data [18,35,39]. The M18 data was used to check if the model is more accurate when we shorten the period between observation and prediction. We used the flattened dense architecture, which was fed with the flattened time series data, to compare the models' capability for capturing temporal relationships with the BiLSTM architecture. Fig. 8(A) depicts the comparison between the deep BiLSTM-based and FFNN-based models. The BiLSTM model was optimized based on multimodal time-series data, and the FFNN models were optimized using BL, M018, and flattened datasets. The BL-FFNN model achieved an average performance of Accuracy= 74.55%, Precision= 75.28%, Recall= 85.94%, and F1-Score= 79.02%. The M18-FFNN model achieved average performance of Accuracy= 72.62%, Precision= 80.83%, Recall= 89.42%, and F1-Score= 72.33%. The Flatten-FFNN model achieved average performance of Accuracy= 73.64%, Precision= 76.44%, Recall= 84.01%, and F1-Score= 73.58%. The time-series-BiLSTM model achieved average performance of Accuracy= 81.22%, Precision= 78.21%, Recall= 82.24%, and F1-Score= 81.23%.

The M18-FFNN model achieved the worst F1- and Accuracy scores, but it achieved the best Recall. The BiLSTM based model achieved the best Accuracy and F1-Scores. Note that, even though the M18-FFNN model achieved the best Recall, the BiLSTM based model achieved the highest F1-Score, which is the mean of both the Recall and Precision. The Flatten-FFNN model did not see any improvement in performance, even though it is based on all of the data over all time steps. The main reason for this is probably due to the huge features created in the flattening process, and the limited capability of the FFNN to learn the temporal features of the data. On the other hand, we observed that M18-FFNN and Flatten-FFNN models achieved comparable performance. It is worth noting that this experiment is based on the DFBL design. Concretely, the four compared models are carry out multiclass classification tasks to predict patient class at M48 using the SoftMax classifier. The learned deep features were tested for the multitask regression optimization design (MRBL). Fig. 8(B) shows the performance of the different models. The deep features of every model are used to optimize the seven regression tasks for CSs at M48, and then these scores are used to predict the patient's class. The BiLSTM based model achieved the highest average results of Accuracy= 78.89%, Precision= 80.79%, Recall= 77.60%, and F1-Score= 79.13%. the BL based model achieved the worst average results of Accuracy= 54.53%, Precision= 65.69%, Recall= 54.38%, and F1-Score= 59.51%. We observe that the BiLSTM based model is also more stable and confident. The average standard

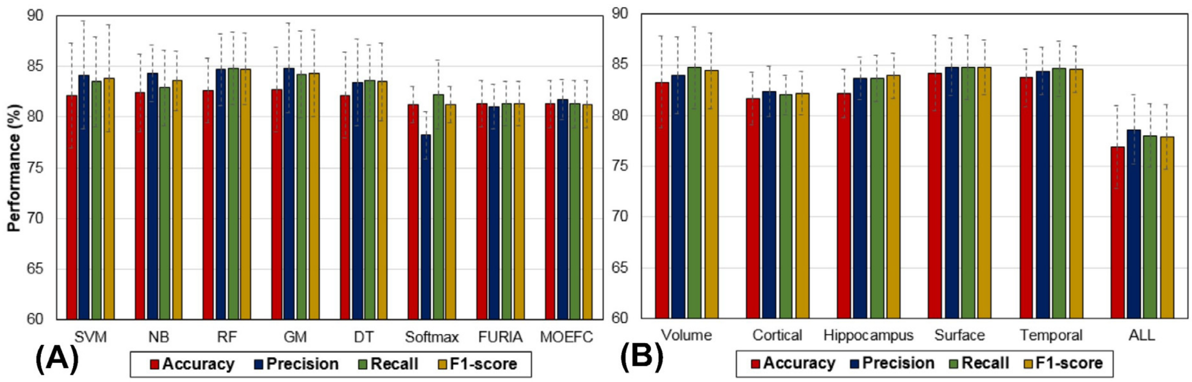


Fig. 9. Comparison between different classifiers based on the learned deep features.

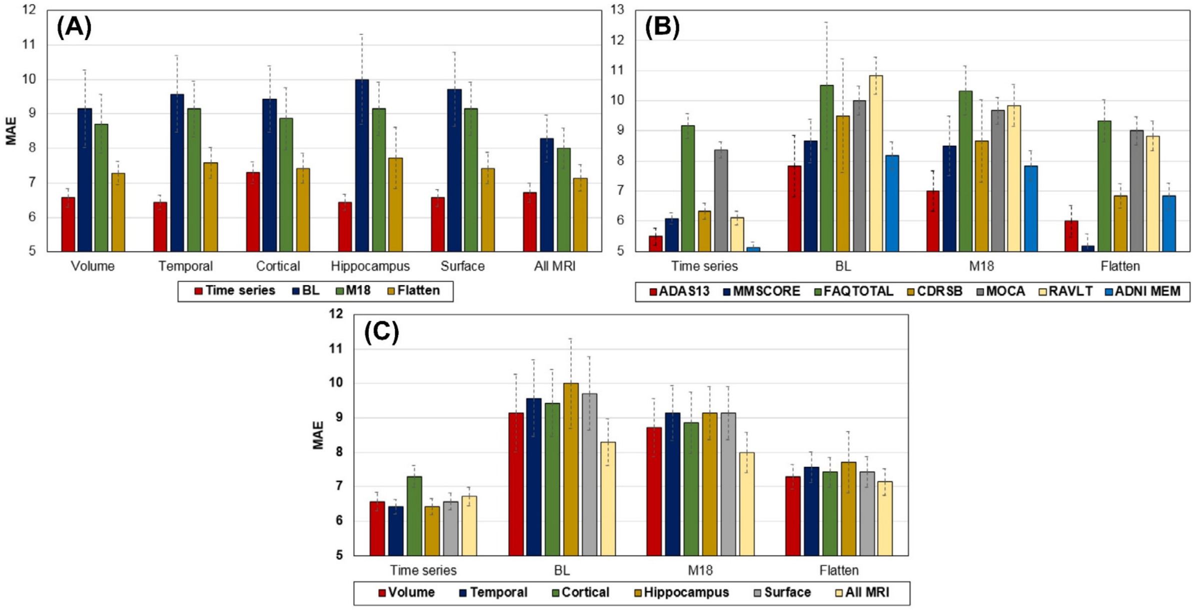


Fig. 10. Performance of the cognitive scores prediction based on different models(i.e., Time series, BL, M18, and Flatten) and used MRI category.

deviations of the models are 8.26%, 8.77%, 8.15%, and 5.84%, for BL, M18, Flatten, and BiLSTM based models, respectively. From these results, we selected the BiLSTM based model to be the main model for learning the deep representations in the time series data. It can be seen that the performance of almost none of the models is high, and this problem is tackled in the next experiments where we integrate the learned deep features with regular ML and fuzzy classifiers.

5.2. Experiment 2: Progression detection based on the DFBL

This experiment investigates the second hypothesis. We explore the effect of using different classifiers with the learned deep features from the deep BiLSTM model. We replaced the deep learning classifier with two types of classifiers, including five regular ML and two fuzzy classifiers. The results are shown in Table 3. This table shows the classification performance of eight classifiers using six different categories of MRI features. Surprisingly, the Softmax classifier did not beat any of the tested classifiers, it achieved results of Accuracy= 81.22%, Precision= 78.21%, Recall= 82.24%, and F1-Score= 81.23%. The lowest performance of the other classifiers was from the MOEFC. It achieved average Accuracy= 81.30%, Precision= 81.72%, Recall= 81.30%, and F1-Score= 81.26%. Fuzzy classifiers commonly have lower performance, but they provide interpretable results in the form of lists of fuzzy rules. FURIA and MOEFC have similar results. The best classifier is RF, which has average Accuracy= 82.63%, Precision= 84.68%, Recall= 84.80%, and F1-Score= 84.73%. Fig. 9(A) illustrates a comparison between the average performance of different models. On average, RF achieved the best results. RF is an ensemble classifier, which is more robust than other individual classifiers such as SVM and NB. We conclude that combining regular ML with deep learning can improve classification task results and beat the performance of the SoftMax classifier.

To explore the role of using different categories of MRI features, we optimize the BiLSTM model based on the used MRI category in combination with other features. Different categories have different results. The list of Surface Area features achieved the highest average results of Accuracy= 84.19%, Precision= 84.77%, Recall= 84.74%, and F1-Score= 84.75%. Using the Temporal region features (see Figure S6 in Supplementary File 2) achieved a comparable result to the Surface Area category, i.e. Accuracy= 83.73%, Precision= 84.38%, Recall= 84.62%, and F1-Score= 84.59%. Fig. 9(B) depicts the average performance of each category regardless of the ML classifier used. As illustrated in Fig. 9(B), using all the MRI features is not a good idea because it achieved the lowest average results of Accuracy= 76.88%, Precision= 78.62%, Recall= 78.03%, and F1-Score= 77.89%.

Table 3
Performance of the DFBL model.

Metric	Modality	SVM	NB	RF	GM	DT	Softmax	FURIA	MOEFC
Accuracy	Volume	86.02 ± 6.85	83.68 ± 5.43	82.87 ± 4.15	86.02 ± 6.85	84.91 ± 6.71	81.03 ± 2.23	81.32 ± 1.79	80.52 ± 2.41
	Cortical	81.75 ± 2.59	79.43 ± 4.49	82.81 ± 2.20	81.75 ± 2.59	83.92 ± 3.53	80.88 ± 0.96	80.88 ± 2.28	81.76 ± 2.18
	Hippocampus	82.92 ± 4.70	78.54 ± 3.44	82.98 ± 2.29	84.04 ± 0.39	82.98 ± 2.29	82.27 ± 1.82	82.27 ± 2.04	81.46 ± 2.04
	Surface	86.08 ± 5.99	88.19 ± 4.52	81.75 ± 2.59	86.08 ± 5.99	82.87 ± 4.15	82.78 ± 2.01	82.63 ± 2.00	83.14 ± 2.52
	Temporal	84.80 ± 2.33	83.86 ± 0.48	84.74 ± 4.75	86.08 ± 4.69	82.87 ± 4.15	81.25 ± 1.62	83.14 ± 2.29	83.07 ± 2.68
	All MRI	71.11 ± 8.42	80.75 ± 4.66	80.64 ± 2.91	72.34 ± 4.36	75.56 ± 4.54	79.13 ± 2.18	77.74 ± 3.37	77.82 ± 2.26
Precision	Volume	86.70 ± 6.59	86.52 ± 2.85	84.70 ± 5.09	86.02 ± 6.85	86.10 ± 6.29	77.01 ± 2.68	81.60 ± 0.02	80.72 ± 0.02
	Cortical	83.48 ± 4.07	78.92 ± 2.57	84.83 ± 3.93	83.48 ± 4.07	85.32 ± 3.03	78.37 ± 2.14	81.40 ± 0.02	81.92 ± 0.02
	Hippocampus	84.34 ± 5.61	84.06 ± 2.42	87.12 ± 1.75	86.67 ± 1.56	85.28 ± 3.65	77.35 ± 1.98	82.59 ± 0.02	81.92 ± 0.02
	Surface	87.20 ± 5.52	89.25 ± 3.36	82.84 ± 2.37	87.20 ± 5.52	83.82 ± 3.92	83.09 ± 1.91	82.85 ± 0.02	81.92 ± 0.02
	Temporal	85.68 ± 3.59	86.03 ± 0.96	86.98 ± 4.26	88.45 ± 3.61	83.82 ± 3.92	77.89 ± 2.35	83.50 ± 0.02	81.92 ± 0.02
	All MRI	77.49 ± 6.72	81.17 ± 4.42	81.60 ± 3.78	77.44 ± 4.66	76.23 ± 4.86	75.53 ± 3.12	77.64 ± 0.03	81.92 ± 0.02
Recall	Volume	88.67 ± 6.31	84.59 ± 5.30	85.81 ± 4.13	88.67 ± 6.31	87.33 ± 6.22	81.09 ± 3.84	81.33 ± 0.02	80.51 ± 0.02
	Cortical	81.78 ± 3.22	79.98 ± 1.42	84.56 ± 3.67	81.78 ± 3.22	84.02 ± 2.79	81.53 ± 3.87	80.88 ± 0.02	81.77 ± 0.02
	Hippocampus	85.13 ± 3.96	79.24 ± 4.95	88.28 ± 3.13	85.88 ± 1.28	84.86 ± 1.97	82.40 ± 2.98	82.27 ± 0.02	81.46 ± 0.02
	Surface	87.47 ± 5.86	88.19 ± 4.52	81.86 ± 2.91	87.47 ± 5.86	83.74 ± 3.13	83.38 ± 3.03	82.64 ± 0.02	83.15 ± 0.03
	Temporal	86.63 ± 3.57	86.11 ± 1.17	85.98 ± 5.69	87.63 ± 4.71	83.74 ± 3.13	80.66 ± 3.76	83.14 ± 0.02	83.07 ± 0.03
	All MRI	71.24 ± 6.60	79.22 ± 5.04	82.32 ± 2.12	73.83 ± 4.43	77.76 ± 3.78	84.35 ± 2.86	77.74 ± 0.03	77.81 ± 0.02
F1-Score	Volume	87.67 ± 6.45	85.54 ± 4.07	85.25 ± 4.52	87.32 ± 6.55	86.71 ± 6.25	81.03 ± 2.23	81.36 ± 0.02	80.55 ± 0.02
	Cortical	82.62 ± 3.64	79.44 ± 1.99	84.69 ± 3.75	82.62 ± 3.64	84.66 ± 2.91	80.88 ± 0.96	81.00 ± 0.02	81.75 ± 0.02
	Hippocampus	84.73 ± 5.73	81.57 ± 3.68	87.69 ± 2.44	86.20 ± 1.42	85.07 ± 2.81	82.27 ± 1.82	82.35 ± 0.02	81.54 ± 0.02
	Surface	87.33 ± 5.69	88.71 ± 2.26	82.34 ± 2.64	87.33 ± 5.69	83.77 ± 3.52	82.78 ± 1.90	82.64 ± 0.02	83.09 ± 0.03
	Temporal	86.15 ± 3.58	86.05 ± 0.98	86.47 ± 4.97	86.83 ± 3.72	83.75 ± 3.21	81.25 ± 1.62	83.14 ± 0.02	83.06 ± 0.03
	All MRI	74.23 ± 6.66	80.14 ± 4.73	81.95 ± 2.91	75.59 ± 4.49	76.98 ± 4.25	79.16 ± 2.16	77.47 ± 0.03	77.58 ± 0.02

Table 4
The MAE results for deep feature extraction based on DL and FFNN models.

Model (time steps)	Modality	ADAS13	MMSCORE	FAQTOTAL	CDRSB	MOCA	RAVLT	ADNI MEM
DL (4 time-steps)	Volume	6.00 ± 0.34	6.00 ± 0.19	09.00 ± 0.42	06.00 ± 0.26	08.00 ± 0.31	06.00 ± 0.20	5.00 ± 0.16
	Temporal	5.00 ± 0.23	6.00 ± 0.11	09.00 ± 0.41	06.00 ± 0.23	08.00 ± 0.21	06.00 ± 0.18	5.00 ± 0.14
	Cortical	5.97 ± 0.31	6.48 ± 0.24	09.97 ± 0.57	06.96 ± 0.27	09.25 ± 0.31	06.61 ± 0.28	5.81 ± 0.23
	Hippocampus	5.00 ± 0.25	6.00 ± 0.17	09.00 ± 0.39	06.00 ± 0.27	08.00 ± 0.27	06.00 ± 0.19	5.00 ± 0.09
	Surface	6.00 ± 0.31	6.00 ± 0.19	09.00 ± 0.36	06.00 ± 0.29	08.00 ± 0.24	06.00 ± 0.22	5.00 ± 0.12
	ALL MRI	5.00 ± 0.23	6.00 ± 0.22	09.00 ± 0.38	07.00 ± 0.26	09.00 ± 0.25	06.00 ± 0.34	5.00 ± 0.28
FFNN (baseline visit)	Volume	7.00 ± 0.85	9.00 ± 0.73	10.00 ± 2.35	09.00 ± 2.56	10.00 ± 0.45	11.00 ± 0.53	8.00 ± 0.41
	Temporal	8.00 ± 1.53	9.00 ± 0.18	11.00 ± 2.91	10.00 ± 1.87	10.00 ± 0.39	11.00 ± 0.51	8.00 ± 0.39
	Cortical	8.00 ± 0.57	9.00 ± 1.37	10.00 ± 0.52	10.00 ± 2.30	10.00 ± 0.59	11.00 ± 0.83	8.00 ± 0.60
	Hippocampus	9.00 ± 1.93	9.00 ± 0.23	12.00 ± 3.46	11.00 ± 1.61	10.00 ± 0.48	11.00 ± 0.75	8.00 ± 0.64
	Surface	8.00 ± 0.78	9.00 ± 0.15	11.00 ± 2.92	10.00 ± 2.25	10.00 ± 0.52	11.00 ± 0.54	9.00 ± 0.37
	ALL MRI	7.00 ± 0.49	7.00 ± 1.66	09.00 ± 0.45	07.00 ± 0.72	10.00 ± 0.47	10.00 ± 0.59	8.00 ± 0.38
FFNN (M18 visit)	Volume	7.00 ± 0.59	9.00 ± 1.33	09.00 ± 0.73	09.00 ± 1.74	09.00 ± 0.35	10.00 ± 0.63	8.00 ± 0.57
	Temporal	7.00 ± 0.93	9.00 ± 0.21	11.00 ± 1.86	09.00 ± 1.34	10.00 ± 0.36	10.00 ± 0.56	8.00 ± 0.37
	Cortical	7.00 ± 0.48	8.00 ± 1.59	10.00 ± 0.51	09.00 ± 1.74	10.00 ± 0.58	10.00 ± 0.78	8.00 ± 0.55
	Hippocampus	7.00 ± 0.84	9.00 ± 0.73	11.00 ± 0.80	09.00 ± 0.91	10.00 ± 0.36	10.00 ± 1.18	8.00 ± 0.63
	Surface	7.00 ± 0.66	9.00 ± 0.63	11.00 ± 0.59	09.00 ± 1.73	10.00 ± 0.74	10.00 ± 0.54	8.00 ± 0.56
	ALL MRI	7.00 ± 0.50	7.00 ± 1.45	10.00 ± 0.36	07.00 ± 0.70	09.00 ± 0.33	09.00 ± 0.46	7.00 ± 0.31
FFNN (flatten 4 time-series)	Volume	6.00 ± 0.41	5.00 ± 0.19	09.00 ± 0.28	06.00 ± 0.23	09.00 ± 0.37	09.00 ± 0.46	7.00 ± 0.47
	Temporal	6.00 ± 0.52	5.00 ± 0.22	10.00 ± 0.54	07.00 ± 0.45	09.00 ± 0.44	09.00 ± 0.36	7.00 ± 0.56
	Cortical	6.00 ± 0.76	5.00 ± 0.36	09.00 ± 0.43	07.00 ± 0.40	09.00 ± 0.32	09.00 ± 0.35	7.00 ± 0.42
	Hippocampus	6.00 ± 0.61	6.00 ± 1.18	10.00 ± 2.09	07.00 ± 0.64	09.00 ± 0.65	09.00 ± 0.42	7.00 ± 0.62
	Surface	6.00 ± 0.40	5.00 ± 0.30	09.00 ± 0.41	07.00 ± 0.34	09.00 ± 0.57	09.00 ± 0.82	7.00 ± 0.33
	ALL MRI	6.00 ± 0.48	5.00 ± 0.24	09.00 ± 0.45	07.00 ± 0.32	09.00 ± 0.48	08.00 ± 0.52	6.00 ± 0.22

5.3. Experiment 3: Multi-tasks regression

In this experiment, we investigated the two related questions of "*is the joint learning of multiple tasks based on the deep BiLSTM model more accurate and stable than using FFNN with BL, M18, and Flatten?*" and "*does the joint learning of multiple regression tasks generate a more robust and stable model than single-task models?*". The first question investigates the role of BiLSTM in multitask modeling, the second one investigates the role of multitask modeling on the performance of regression tasks based on the BiLSTM model. Regarding the first question, Table 4 shows the results based on the MAE for the seven regression tasks. The results are reported for the six MRI categories. We investigated four models: deep BiLSTM with time series, and FFNN with BL, M18, and Flatten data. It is clear from the table that BiLSTM has a lower error rate for every regression task compared to the FFNN models. The detailed performance of every regression task for each model is illustrated in Figure S6 in Supplementary file 2. The box plots of Figure S2, assert that BiLSTM based SCs are more accurate and less noisy than FFNN based scores. The average performance of these four models is illustrated in Fig. 10(A). The BiLSTM model is more accurate than all other FFNN models because it achieved the lowest error rates in all MRI categories. The BiLSTM model achieved MAE values of 6.5714, 6.4286, 7.2929, 6.4286, 6.5714, and 6.7143 for Volume, Temporal, Cortical, Hippocampus, Surface, and All MRI categories, respectively. The average MAE rates for the BiLSTM, BL-FFNN, M18-FFNN, and Flatten-FFNN are 6.6679, 9.3571, 8.8333, and 7.4286. More importantly, the BiLSTM model is more stable than the other models. The average standard deviations for the BiLSTM, BL-FFNN, M18-FFNN, and Flatten-FFNN are 0.2598, 1.0435, 0.7809, and 0.4914. As noticed the Flatten-FFNN model is more stable than other FFNN models. This could

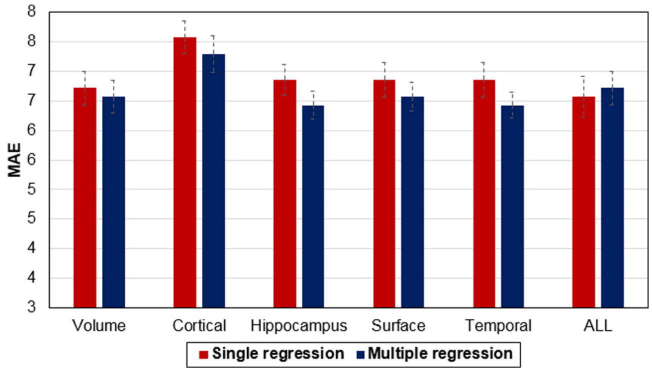


Fig. 11. A comparison between single task and multitask regression.

be a result of using all flattened time-series features. Fig. 10(B) explores the best performing CSs for every design model. It is of note that the BiLSTM model achieved the lowest MAE rates for the seven tasks. ADNI MEM had the best cognitive score (MAE= 5.135) in the BiLSTM model, and FAQTOTAL had the worst (MAE= 9.162). The average MAE rates for ADAS13, MMSCORE, FAQTOTAL, CDRSB, MOCA, RAVLT, ADNI MEM are 6.582, 7.103, 9.832, 7.832, 9.26, 8.9, and 6.992, respectively. On average ADAS 13 achieved the best performance and FAQTOTAL the worst. As shown in Fig. 10(B), the BL-FFNN model achieved the worst results for all CSs regression tasks. It has MAE rates of 7.833, 8.667, 10.5, 9.5, 10, 10.833, and 8.167 for ADAS13, MMSCORE, FAQTOTAL, CDRSB, MOCA, RAVLT, ADNI MEM, respectively. This is obvious because the learning of the joint features for seven tasks and the optimization of such complex objective functions is too difficult to be learned in a single time step, especially using the BL step. Fig. 10(C) explores the role of different MRI categories on the performance of the BILSTM and FFNN models. Again, the BiLSTM model achieved the best results for all MRI categories. For the BiLSTM model, the best MRI feature groups are the Hippocampus and Temporal, i.e. MAE= 6.429, and using All MRI features achieved the highest MAE of 6.714. On average, the MAE for the Volume, Temporal, Cortical, Hippocampus, Surface, and All MRI are 7.929, 8.179, 8.252, 8.321, 8.214, and 7.536, respectively. The BL-FFNN achieved the highest MAE rates of 9.143, 9.571, 9.429, 10, 9.714, and 8.286 for Volume, Temporal, Cortical, Hippocampus, Surface, and All MRI. To conclude, the previous results asserted the important role of the deep BiLSTM model to optimize a complex multitask objective function. To answer the second question, we have compared the multitask model with 42 single-task models (7 CSs × 6 datasets of different subsets of MRI data). Fig. 11 compares the performance of different data fusions by averaging the results of the seven tasks. Multitask modeling achieved statistically significantly better results (P= 0.01, Student t-test). In addition, the multitask models are more stable because the models have lower variance than the single-task models.

5.4. Experiment 4: Progression detection based on the MRBL

This experiment investigates the question "*Is it possible to predict the future values of a set of cognitive scores and use them to predict AD progression?*" We predicted seven critical CSs at M48 and used these scores to classify the patient as CN vs. MCI vs. AD. We depended on the learned CSs from BiLSTM, as discussed in Experiment 3. The predicted scores are combined with age, gender, and education features from background data. Note that the added features are not affected by the time, so it is safe to integrate them with predicted features. The prepared dataset was used to optimize five regular ML classifiers, Softmax, and two fuzzy classifiers. Table 5 shows the results for all classifiers. We checked the role of different MRI feature types. Fig. 12 illustrates the average performance with respect to the ML model (Fig. 12(A)) and the MRI category (Fig. 12(B)). As shown in Fig. 12(A), the RF ensemble classifier achieved the best average results of Accuracy= 80.30%, Precision= 81.65%, Recall= 80.80%, and F1-score= 81.19%. FURIA achieved the worst results of Accuracy= 68.90%, Precision= 74.81%, Recall= 68.90%, and F1-score= 68.39%. However, MOEFC achieved a comparable average result to other ML models, i.e. Accuracy= 79.08%, Precision= 79.06%, Recall= 81.06%, and F1-score= 78.87%. The Softmax classifier achieved average results of Accuracy= 78.89%, Precision= 80.79%, Recall= 77.60%, and F1-score= 79.13%. MOEFC is the most stable because it has the lowest standard deviation of all performance metrics. i.e. Accuracy= 2.06%, Precision= 2.17%, Recall= 2.17%, and F1-score= 2%. As a result, integrating regular ML models with an accurate DL model could improve performance. DL is better at learning deep feature representations from complex data structures like time series. Furthermore, DL models can jointly learn multiple tasks to produce more stable models. However, adding a DL classifier to the extracted features usually requires a large amount of data. Regular ML classifiers could help in this situation because they can be optimized using smaller amounts of data. Fig. 12(B) compares the role of MRI categories in the MRBL design. Using All MRI features achieved the worst average results of Accuracy= 68.34%, Precision= 69.31%, Recall= 68.32%, and F1-score= 67.66%. The Surface and Volume categories achieved the highest performance of Accuracy= 80.60%, Precision= 82.72%, Recall= 81.28%, and F1-score= 81.59%.

5.5. Comparison of DFBL and MRBL architectures

In this section, we compare the overall performance of MRBL and DFBL design architectures. Fig. 13(A) shows a comparison between eight models based on the FFNN and BiLSTM architectures for deep feature learning. To explore the role of regular ML models as classifiers, we compare the best ML (i.e. RF) and Softmax models. All ML models based on BiLSTM features are more accurate than other models based on FFNN. These models achieved the best results, either using the RF or Softmax classifiers. Using the BiLSTM deep feature architecture: (1) the DFBL-RF achieved the best results of Accuracy= 82.63±1.803%, Precision= 84.68±2.363%, Recall= 84.80±3.390%, and F1-score= 84.73±1.782%; (2) MRBL-RF achieved the second-best results of Accuracy= 80.30±3.983%, Precision= 81.65±2.953%, Recall= 80.80±3.962%, and F1-score= 81.19±3.692%. Besides, these models are more stable than other models. Using the Softmax classifier with BiLSTM deep features, the DFBL-Softmax architecture achieved better results (Accuracy= 81.22±3.148%, Precision= 78.21±3.530%, Recall= 82.24±3.608%, and F1-score= 81.23±3.538%) than the MRBL-Softmax architecture (Accuracy= 78.89±4.955%, Precision= 80.79±5.703%, Recall= 77.60±6.578%, and F1-score= 79.13±6.138%). It is of note that using the DFBL to extract deep features from time-series data using the BiLSTM model and feeding the extracted deep features to an accurate classifier like RF results in a more accurate and more stable architecture. However, the resulting model is not interpretable because it is based on deep features. On the other hand, optimizing the RF classifier using the predicted CSs at M48 did not result in a highly accurate model, but the generated model is highly interpretable. This model is based on ten medically interpretable features. The visualization of the resulting RF can be easily understood by medical experts [78], and the collection of rules from the RF's trees can be extracted and summarized using tools like inTrees [79]. In addition, the fuzzy

Table 5
The performance of the MRBL model.

Metric	Modality	SVM	NB	RF	GM	DT	Softmax	FURIA	MOEFC
Accuracy	Volume	82.63 ± 2.92	81.68 ± 4.22	83.47 ± 2.56	81.68 ± 4.22	83.68 ± 2.10	83.53 ± 4.23	67.16 ± 6.58	80.59 ± 1.89
	Cortical	72.42 ± 5.85	71.42 ± 5.72	82.89 ± 6.57	62.00 ± 7.61	77.53 ± 4.62	78.32 ± 2.64	67.74 ± 3.30	77.96 ± 2.52
	Hippocampus	79.63 ± 3.28	80.42 ± 2.18	84.95 ± 9.28	80.42 ± 2.18	82.53 ± 2.31	80.63 ± 6.66	65.84 ± 4.54	80.81 ± 1.92
	Surface	78.53 ± 6.91	81.47 ± 2.54	84.95 ± 6.98	80.63 ± 2.05	80.53 ± 2.76	81.42 ± 2.95	75.33 ± 2.98	82.34 ± 2.06
	Temporal	78.58 ± 4.26	79.58 ± 3.77	82.84 ± 8.90	79.63 ± 3.28	79.58 ± 0.58	79.63 ± 3.28	67.74 ± 4.86	80.81 ± 1.92
	All MRI	69.37 ± 5.21	75.26 ± 2.16	62.74 ± 3.44	64.84 ± 5.42	63.16 ± 3.87	69.79 ± 4.14	69.57 ± 2.26	71.98 ± 2.04
Precision	Volume	84.82 ± 4.49	86.29 ± 0.55	86.18 ± 2.03	85.12 ± 2.45	85.03 ± 4.35	86.65 ± 2.49	74.04 ± 0.05	80.50 ± 0.02
	Cortical	74.87 ± 6.04	74.44 ± 4.24	85.17 ± 7.23	74.21 ± 5.25	77.56 ± 4.20	81.05 ± 5.13	74.32 ± 0.02	78.03 ± 0.03
	Hippocampus	80.17 ± 2.74	81.35 ± 2.92	85.65 ± 8.52	81.35 ± 2.92	83.14 ± 2.48	83.88 ± 4.79	73.08 ± 0.03	80.71 ± 0.02
	Surface	78.50 ± 6.33	82.83 ± 3.12	87.00 ± 6.61	82.08 ± 2.68	81.42 ± 1.33	82.62 ± 3.00	78.1 ± 0.02	82.26 ± 0.02
	Temporal	80.68 ± 3.24	81.65 ± 3.38	84.37 ± 7.99	81.21 ± 4.09	80.63 ± 3.15	80.94 ± 0.91	74.26 ± 0.03	80.71 ± 0.02
	All MRI	76.03 ± 6.53	79.00 ± 3.09	61.52 ± 1.84	60.18 ± 3.58	60.95 ± 3.14	69.59 ± 6.40	75.08 ± 0.03	72.15 ± 0.02
Recall	Volume	83.90 ± 3.55	82.41 ± 4.58	84.66 ± 1.59	82.00 ± 4.70	84.84 ± 3.28	84.75 ± 3.31	67.16 ± 0.07	80.58 ± 0.02
	Cortical	74.06 ± 7.11	72.57 ± 6.19	82.66 ± 6.47	62.74 ± 7.81	79.08 ± 4.27	78.52 ± 3.68	67.75 ± 0.03	77.97 ± 0.03
	Hippocampus	79.99 ± 3.78	81.13 ± 3.00	86.50 ± 9.39	81.13 ± 3.00	83.35 ± 2.63	80.29 ± 6.67	65.84 ± 0.05	80.82 ± 0.02
	Surface	80.15 ± 5.54	82.26 ± 3.58	84.71 ± 7.80	81.59 ± 2.93	81.95 ± 3.58	81.88 ± 3.40	75.33 ± 0.03	82.33 ± 0.02
	Temporal	79.94 ± 5.01	80.25 ± 3.91	83.83 ± 8.59	80.02 ± 4.06	80.52 ± 4.71	79.94 ± 1.81	67.74 ± 0.05	82.33 ± 0.02
	All MRI	69.60 ± 4.99	75.59 ± 2.31	62.41 ± 5.63	63.91 ± 6.19	62.86 ± 5.40	60.24 ± 4.90	69.58 ± 0.02	82.33 ± 0.02
F1-Score	Volume	84.36 ± 3.11	84.31 ± 2.54	85.41 ± 1.81	83.53 ± 3.57	84.93 ± 3.81	85.69 ± 2.80	65.80 ± 0.09	80.27 ± 0.02
	Cortical	74.46 ± 6.57	73.50 ± 5.21	83.90 ± 6.85	67.99 ± 6.53	78.31 ± 4.23	79.76 ± 4.41	67.64 ± 0.04	77.78 ± 0.02
	Hippocampus	80.08 ± 3.51	81.24 ± 2.64	86.07 ± 8.95	81.24 ± 2.96	83.24 ± 2.55	82.05 ± 5.73	64.67 ± 0.06	80.57 ± 0.02
	Surface	79.32 ± 3.21	82.53 ± 4.33	85.84 ± 7.20	81.83 ± 2.80	81.68 ± 2.45	82.25 ± 3.20	75.64 ± 0.03	82.03 ± 0.02
	Temporal	80.31 ± 4.25	80.94 ± 3.64	84.10 ± 8.29	80.61 ± 4.07	80.57 ± 3.66	80.44 ± 1.36	67.13 ± 0.06	80.57 ± 0.02
	All MRI	72.33 ± 5.76	77.25 ± 2.69	61.82 ± 3.73	61.98 ± 4.88	61.89 ± 4.27	64.57 ± 5.65	69.43 ± 0.03	71.99 ± 0.02

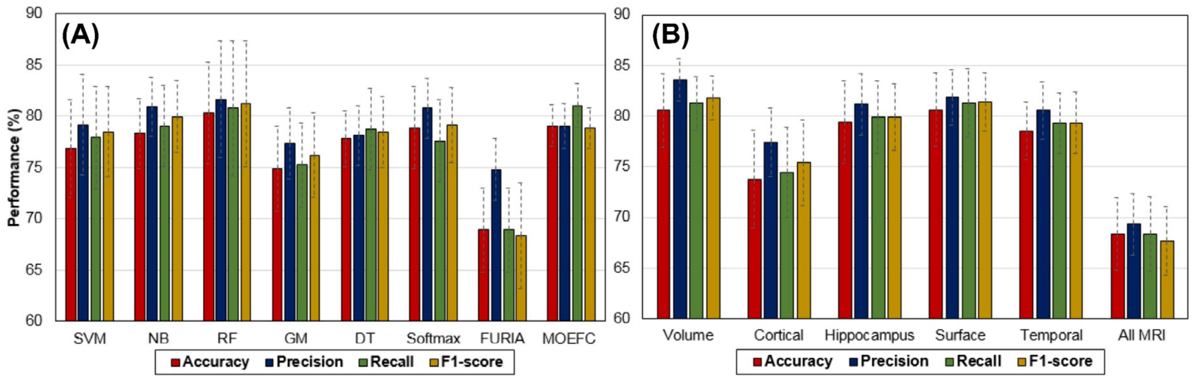


Fig. 12. Performance of different classifiers based on the MRBL architecture.

classifier in the MRBL can generate a medically acceptable model. This issue is discussed in the next section. The FFNN based models are less accurate and noisier. The best Accuracy of an MRBL-based model is the Flatten-FFNN-Softmax model (64.87± 9.028%) and the best of a DFBL-based model is the BL-FFNN-Softmax model (74.55± 1.76%). The best Precision of an MRBL-based model is the Flatten-FFNN-Softmax model (69.33±6.99%) and of a DFBL-based model is the M18-FFNN-Softmax model (80.82±1.70%). The best Recall of an MRBL-based model is the Flatten-FFNN-Softmax model (66.86±8.98%) and of a DFBL-based model is the M18-FFNN-Softmax model (89.42±3.53%). The best F1-Score of an MRBL-based model is the Flatten-FFNN-Softmax model (67.65±7.60%) and of a DFBL-based model is the BL-FFNN-Softmax model (79.02±1.50%). The Flatten-FFNN-Softmax model achieved better overall performance compared to other FFNN-based models. However, it has much fewer results compared to the BiLSTM-based models including RF and Softmax. Fig. 13(B) illustrates a comprehensive comparison between the different design architectures for the performance of all MRI categories from both FFNN and BiLSTM architectures. It can be seen that the BiLSTM based models achieved the best average results for either MRBL or DFBL based on the Surface area category. Please note that these averages are calculated from the performance of the five regular ML, the Softmax, and the two fuzzy models. The MRBL-based design has an average performance of Accuracy= 80.65±3.72%, Precision= 81.85±2.83%, Recall= 81.28±3.17%, and F1-score= 81.39±2.72%. The DFBL-based design has the average performance of Accuracy= 84.19±3.65%, Precision= 84.77±2.76%, Recall= 84.74±3.36%, and F1-score= 84.75±2.90%. The best FFNN-based model is FFNN-Hippocampus-DFBL, where the Hippocampus category is used with the DFBL design. It achieved average performance of Accuracy= 74.32±5.06%, Precision= 79.32±3.76%, Recall= 80.02±5.70%, and F1-score= 76.46±3.75%. All of these results are less accurate than the BiLSTM-based architecture, which highlights the crucial role of time series analysis in deep BiLSTM models.

5.6. Experiment 5: Explainability of progression detection decision

The previous list of experiments concentrated on the performance of the resulting model. However, in the medical domain, accuracy is not the only measure of acceptable systems. The interpretability of the model is also a crucial requirement because physicians are expected to sufficiently understand and trust the model before they start using it [80]. In ML, there is a trade-off between performance (e.g. accuracy) and interpretability. Transparent models that are considered interpretable, such as LR, KNN, DT, and FL models, often perform worse than black-box models, such as DL, SVM, and RF. Medical experts prefer a

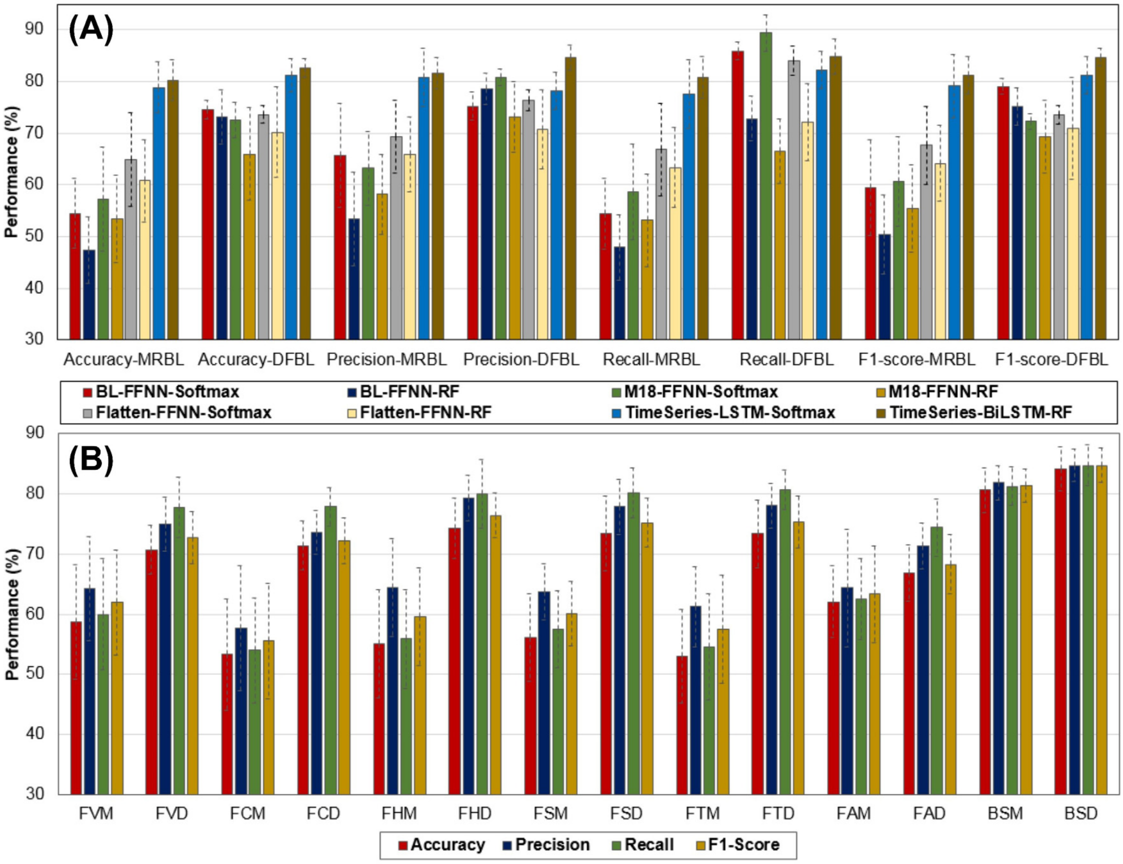
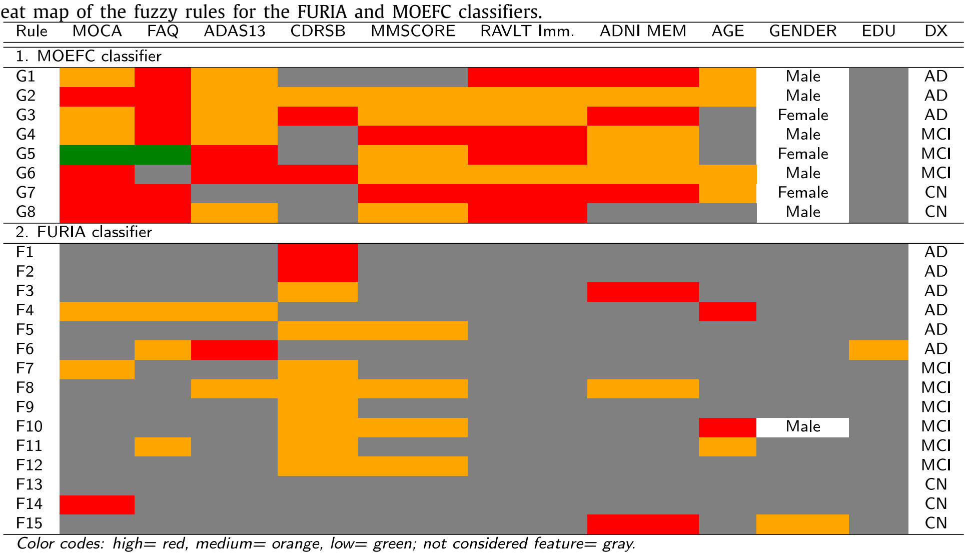


Fig. 13. Comparison between the best models from DFBL and MRBL architectures. FVM (FFNN-Volume-MRBL), FVD (FFNN-Volume-DFBL), FCM (FFNN-Cortical-MRBL), FCD (FFNN-Cortical-DFBL), FHM (FFNN-Hippocampus-MRBL), FHD (FFNN-Hippocampus-DFBL), FSM (FFNN-Surface-MRBL), FSD (FFNN-Surface-DFBL), FTM (FFNN-Temporal-MRBL), FTD (FFNN-Temporal-DFBL), FAM (FFNN-All MRI-MRBL), FAD (FFNN-All MRI-DFBL), BSM (BiLSTM-Surface-MRBL), and BSD (BiLSTM-Surface-DFBL).

Table 6
Heat map of the fuzzy rules for the FURIA and MOEFC classifiers.



Rule	MOCA	FAQ	ADAS13	CDRSB	MMSCORE	RAVLT Imm.	ADNI MEM	AGE	GENDER	EDU	DX
1. MOEFC classifier											
G1	🟧	🟥	🟧	▨	▨	🟥	🟥	🟧	Male	▨	AD
G2	🟥	🟥	🟧	🟧	🟧	🟧	🟧	🟧	Male	▨	AD
G3	🟧	🟥	🟧	🟥	🟧	🟧	🟥	▨	Female	▨	AD
G4	🟧	🟥	🟧	▨	🟥	🟥	🟧	▨	Male	▨	MCI
G5	🟩	🟩	🟥	▨	🟧	🟥	🟧	▨	Female	▨	MCI
G6	🟥	▨	🟥	🟥	🟧	🟧	🟧	🟧	Male	▨	MCI
G7	🟥	🟥	▨	▨	🟥	🟥	🟥	🟧	Female	▨	CN
G8	🟥	🟥	🟧	▨	🟧	🟥	▨	▨	Male	▨	CN
2. FURIA classifier											
F1	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨	AD
F2	▨	▨	▨	🟥	▨	▨	▨	▨	▨	▨	AD
F3	▨	▨	▨	🟧	▨	▨	🟥	▨	▨	▨	AD
F4	🟧	🟧	🟧	▨	▨	▨	▨	🟥	▨	▨	AD
F5	▨	▨	▨	🟧	🟧	▨	▨	▨	▨	▨	AD
F6	▨	🟧	🟥	▨	▨	▨	▨	▨	▨	🟧	AD
F7	🟧	▨	▨	🟧	▨	▨	▨	▨	▨	▨	MCI
F8	▨	▨	🟧	🟧	🟧	▨	🟧	▨	▨	▨	MCI
F9	▨	▨	▨	🟧	▨	▨	▨	▨	▨	▨	MCI
F10	▨	▨	▨	🟧	🟧	▨	▨	🟥	Male	▨	MCI
F11	▨	🟧	▨	🟧	▨	▨	▨	🟧	▨	▨	MCI
F12	▨	▨	▨	🟧	🟧	▨	▨	▨	▨	▨	MCI
F13	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	CN
F14	🟥	▨	▨	▨	▨	▨	▨	▨	▨	▨	CN
F15	▨	▨	▨	▨	▨	▨	🟥	▨	🟧	▨	CN

Color codes: high= red, medium= orange, low= green; not considered feature= gray.

Color codes: high = red, medium = orange, low = green; not considered feature = gray.

balanced model that is as accurate and interpretable as possible. For example, in Caruana et al. [81], LR was chosen by domain experts over FFNN because of interpretability concerns. Although FFNN achieved a significantly higher ROC score than the LR, the experts considered it too risky to deploy a black-box model like FFNN for decision making with real patients. The LR, on the other hand, though less accurate, provides physicians with a transparent and interpretable model, which can facilitate the investigation of problematic patterns in data. *It is worth noting that there is no previous study on the interpretability of hybrid multimodal multitask DL models in the AD domain*. According to the previous experiments above, the DFBL with RF achieved the best results. However, all models based on the DFBL are not interpretable because they are based on deep features extracted from the BiLSTM model. As a result, to achieve interpretability, we should concentrate on the MRBL architecture. It is possible to calculate the importance of the learned features with MRBL (Supplementary File 2, Figure S7). We have calculated the importance of the eight learned CSs and integrated three demographics using many techniques including information gain, Gini index, gain ratio, correlation, feature importance of the XGBoost classifier, and permutation importance using RF. All methods agreed that CDRSB is the best feature to predict AD progression, followed by FAQ. The least important feature is gender. Using the level of importance, domain experts can understand why a model has taken specific decision for a specific patient. Further, we used many ML models with the MRBL design and the most interpretable models are the DT, FURIA, and MOEFC. All these models can formulate their decision behavior as a list of rules. Nevertheless, the performance of the RF model is better than all explainable models, see Table 5. The best accuracy achieved by the RF model is 84.95% using hippocampus and surface area feature spaces. The best accuracy of the DT is 83.68% using the volume feature set. Regarding FURIA and MOEFC, the best accuracies are 75.33% and 82.34%, respectively, using the surface area feature space. It should be noted that DT and MOEFC achieved acceptable results compared to RF. These models are expected to provide an accurate explanation in connection with the RF decisions. As MOEFC is based on genetic optimization and Gaussian membership functions, it achieved better results than FURIA. Simple rule-based models are commonly considered to be interpretable [80]. The MOEFC generated eight rules (3 for AD, 3 for MCI, and 2 for CN), FURIA generated 15 rules (6 for AD, 6 for MCI, and 3 for CN), and DT generated 21 rules (7 for AD, 11 for MCI, 2 for CN). The full list of rules can be found in Supplementary Table S2, Table S3, and Table S4. Table 6 illustrates the utilized features for the rule-based FURIA and MOEFC. On average four features are used in the antecedent of the FURIA rules and nine features are utilized by the MOEFC model. The Education feature was completely ignored by MOEFC, while the RAVLT features was ignored by FURIA. Please note that we could not add the DT rules in Table 6 because the decision boundaries of these rules are crisp.

5.7. Comparison with related studies

The methods proposed in this study show advances in both deep learning and AD progression detection. (1) Regarding the DL proposal, first, we explored the role of late fusion of multimodal multivariate time series data using BiLSTM models to improve the performance of DL models. Second, we investigated the role of hybrid modeling of DL feature extraction and regular ML classifiers. According to our results, we discovered that replacing the output layer of DL models with more accurate classifiers such as RF improved the performance of the overall model. We checked the performance of many regular ML classifiers for performing the classification step of the deep learning model. In addition, we investigated the role of two popular fuzzy classifiers to do the same task. Third, we explored the performance of multitask modeling to generate more accurate and stable DL models. We jointly trained models to simultaneously predict seven CSs at M48. (2) Regarding AD progression detection, first, we utilized five popular time-series modalities in addition to the baseline data to predict AD progression within 2.5 years from last visit. Each time series modality was separately learned by a deep BiLSTM model and the resulting deep features of these heterogeneous modalities were fused using another deep BiLSTM model. The resulting features were further fused with the learned features from the baseline data. AD progression detection was explored using DFBL and MRBL techniques. Second, we studied the role of different MRI feature sets to enhance the performance of every model. The discriminating capabilities of the volume, cortical thickness, hippocampus region, surface areas, temporal region, and all MRI features were explored for both the DFBL and MRBL techniques. Third, we utilized the feature importance, fuzzy classifiers, and DT to interpret the decisions of DL models. To the best of our knowledge, the proposed models have not been explored in the AD domain before. In addition, we achieved accurate, stable, and interpretable results. El-Sappagh et al. [83] proposed a DL model for AD progression detection based on a multimodal multitask paradigm. The current proposal can be considered as an extension to that by El-Sappagh et al. [83]. As an example of the extensions, in the MRBL, we predicted more CSs using multitask modeling and utilized a different fusion mechanism by using multiple layers of decision fusion based on BiLSTM. Table 7 shows a comparison of models from the most recent literature on AD progression detection. As can be clearly seen, all previous studies have concentrated on a single narrow area for their evaluations. For example, [6,42,74,82,84] modeled AD progression as a simple binary classification task. The majority of the AD studies are mainly based on the analysis of the MRI data [26,30,57,65,74,84,85,87,88,90,92,96]. The deep learning models proposed for dealing with baseline MRI data are mostly based on a CNN architecture [30,42,84,92]. Time-series data analysis is not popular in the AD domain because there are not enough data and the data has long time steps in between. Lee et al. [6] proposed a binary classifier based on the early fusion of demographics, MRI, CSs, and CSF data. This study collected four time-steps and achieved 81% accuracy using a GRU model. Cui et al. [74] utilized a stacked CNN-BGRU to build a binary classifier based on MRI data only. The model achieved an accuracy of 91.33% for AD vs. NC and 71.71% for pMCI vs. sMCI tasks. El-Sappagh et al. [83] utilized a longer time series of 15 time-steps; however, the authors did not use a gap between the last observed data and the prediction time step. Platero and Tobar [89] used six time-steps and Hong et al. [57] used ten time-steps. Most time-series data analysis utilized RNN models [6,57,74,83,89]. Due to space restrictions, we will not discuss the details of every study, Table 7 provides a nice comparison of nine metrics and provides the most critical features from every study. The proposed model can be used as a starting point to build a clinical decision support system for Alzheimer's disease progression detection. The model is more accurate than state-of-the-art studies. In addition, the model could provide the explainability of the suggested decisions. Moreover, the proposed model is medically more intuitive than the current literature because it is based on the multimodal longitudinal data of different critical modalities including MRI, PET, neuropsychological battery, cognitive subscores, and neuropathology. Several baseline features have been combined in the model to support the model providing more robust decisions. The proposed model can provide customized and individualized decisions based on the patient's complete profile. Although our model provides an advanced point in the fields of AD management and deep learning, the model still needs further improvements to be used in real environments. The interoperability of the clinical decision support system with the hospital information systems is complex and needs further investigation. The more comprehensive explainability techniques model decisions such as ontology and case-based reasoning, neuroimage visualization, and knowledge-based systems should be considered in future work.

Table 7
A comparison with the literature studies.

Study	Subjects	Tasks (classes)	Reg. tasks	Modality	Fusion	Time steps	Performance (%)	Method
[6]	1618	1 (2)	–	D, MRI, CSs, CSF	Early	4	Accuracy (81%)	GRU
[26]	896	1 (2)	–	MRI	–	–	Accuracy (CN/MCI: 78.8%)	Gaussian process
[30]	1984	1 (4)	4	MRI	Late	–	Accuracy (4-classes: 51.8%), RMSE (CDRSB, ADAS 11, ADAS 13, MMSE: 1.666, 6.2, 8.537, 2.373)	CNN
[42]	785	2 (2)	–	MRI, D, N, and APOe4	Early	–	Accuracy (sMCI/pMCI: 86%, CN/AD: 100%)	CNN
[57]	1105	1 (2)	–	MRI	–	10	Accuracy (AD/CN: 93.5%, CN/MCI: 69.7%, MCI/AD: 79.8%, 3-classes: 77.7%)	LSTM
[58]	488	1 (3)	1	CSs, PET, MRI, CSF	Early	–	Accuracy (3-classes: 83.0%), R2 = 0.874	SVM, Ridge
[65]	485	1 (3)	–	MRI	–	–	Accuracy (CN/AD: 0.94%, MCI/AD: 87%, CN/MCI: 95%, 3-classes: 82%)	LR, KNN, SVM, DT, RF
[82]	237	1 (2)	–	10 modalities	Early	–	Accuracy (73%)	SVM
[74]	830	2 (2)	–	MRI	–	6	Accuracy (AD/NC: 91.33%, pMCI/sMCI: 71.71%)	Stacked CNN-BGRU
[83]	1536	1 (4)	4	MRI, PET, CSs, N	Late	15	Accuracy (4-classes: 92.62%), MAE(FAQ, ADAS, CDR, MMSE: 0.107, 0.076, 0.075, and 0.085)	Stacked CNN-BiLSTM
[84]	694	1 (2)	–	MRI	–	–	Accuracy (CN/AD: 86.60%, CN/pMCI: 77.37%, CN/sMCI: 63.04%, AD/pMCI: 60.97%, sMCI/AD: 75.06%)	CNN
[85]	458	1 (4)	4	MRI	–	–	Accuracy (CN/AD: 93.01%, pMCI/sMCI: 75.00%)	SVM
[86]	906	1 (3)	–	MRI, AV45 PET	Early	–	Accuracy (CN/sMCI: 79.25%, 3-classes:75.28%, sMCI/pMCI/AD: 67.69%)	GDCA
[87]	449	1 (2)	–	MRI	Late fusion	–	Accuracy (CN/AD: 88.9%, CN/MCI: 76.2%)	CNN, DenseNet
[88]	400	1 (4)	–	MRI	Early	–	Accuracy (4-classes: 61.9%)	RF
[89]	321	1 (2)	–	MRI and N	Early	6	Accuracy (MCI/AD: 78% baseline, 85% time series)	LDA
[90]	828	1 (4)	–	MRI	–	–	Accuracy(4-classes: 83.01%)	ResNet
[91]	1051	1 (2)	–	FDG-PET	–	–	Accuracy (NC/AD: 93.58%, sMCI/pMCI: 81.55%, sMCI/pMCI: 82.51%)	DNN
[92]	229	1 (2)	–	MRI	–	–	Accuracy (sMCI/pMCI: 75%, CN/AD: 99%,	CNN
[93]	756	1 (2)	–	MRI, CSs	Early	–	Accuracy(sMCI/pMCI: 87%)	Naïve Bayes
[94]	475	1 (3)	–	MRI	–	–	Accuracy (MCI/AD: 81.3%)	two stage classifiers
[95]	290	1 (2)	–	CSF, CSs, MRI	Early	–	Accuracy (sMCI/pMCI: 86.4%)	SVM
[96]	509	1 (2)	–	MRI	–	–	Accuracy (CN/ AD: 84%, CN/pMCI: 79%, sMCI/pMCI: 62%)	Ensemble CNN
[97]	1029	1 (4)	–	C, CSs, MH	–	–	Accuracy (3-classes: 87.69%, 4-classes: 83.68%)	LR, KNN, SVM, DT, RF
This work	**1371**	**2 (3)**	**7**	**MRI, PET, CSs, N, NP, D**	**Late**	**4**	**Accuracy (3-classes: 84.95% (MRBL), 86.08% (DFBL))**	**Hybrid/ multitask BiLSTM**

4-classes: CN/sMCI/pMCI/AD, 3-classes: CN/MCI/AD, D: Demographics, N: Neuropsychological, NP: Neuropathology, MH: Medication History, C: comorbidities.

6. Conclusion

In this paper, we proposed and compared two BiLSTM-based deep learning models (i.e., MRBL and DFBL models) for AD progression detection. The models are based on the fusion of multimodal time series data collected from the ADNI dataset. The integrated time series data includes neuroimaging (i.e. MRI, PET), cognitive subscores, neuropathology, and neuropsychological battery data. In addition, we explored the role of different MRI features to improve the prediction performance. The extracted deep features from the deep BiLSTM model were used to train the MRBL and DFBL models. The MRBL model is based on the fusion of the learned deep features and the patient demographics at baseline. It is based on a multitask modeling paradigm, where the model has been optimized to learn seven medically related regression tasks. Each task is responsible for predicting a specific cognitive score at month 48, i.e. 2.5 years after final observation. The collected scores at M48 are used to predict the AD progression at that time. We compared multiple models (SoftMax, FURIA, MOEFC, RF, DT, SVM, GM, and NB) to predict AD progression based on these collected scores. The RF ensemble classifier achieved the best average results of Accuracy= 80.30%, Precision= 81.65%, Recall= 80.80%, and F1-score= 81.19%. The predicted scores at M48 can be used to build an explainable model, which facilitates domain experts to understand why the model has taken a specific decision. We optimized DT and fuzzy classifiers (FURIA and MOEFC) to predict AD progression. These models achieved good performance, and they are interpretable at the same time. On the other hand, the DFBL model is a hybrid BiLSTM-ML model, where BiLSTM is used to extract deep features from the multimodal time-series data, and the ML classifier is optimized to perform the classification task. The architecture was compared when using the regular SoftMax classifier to when using popular ML classifiers like SVM and RF. RF achieved better average performance (i.e. Accuracy= 82.63%, Precision= 84.68%, Recall= 84.80%, and F1-Score= 84.73%) than SoftMax. These results highlight a possible role for ML models to work as classifiers that are integrated with deep learning architectures. Although the DFBL architecture is more accurate than the MRBL architecture, the latter is more interpretable. Accordingly, the MRBL architecture is more trustful and acceptable by medical experts who usually prefer balanced models (i.e., models which achieve acceptable performance but provide users also with explainable outcome). The impact of fine-tuning the proposed models as well as evaluating the models run time complexity will be explored in the future work.

CRediT authorship contribution statement

Tamer Abuhmed: Conceptualization, Methodology, Software, Writing. **Shaker El-Sappagh:** Conceptualization, Methodology, Writing. **Jose M. Alonso:** Data curation, Writing - Original draft.

Acknowledgments

Dr. Jose M. Alonso is Ramon y Cajal Researcher (RYC-2016-19802), and its research is supported by the Spanish Ministry of Science, Innovation, and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, ED431G/08, and ED431G2019/04), with all grants co-funded by the European Regional Development Fund (ERDF/FEDER program). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2016R1D1A1A03934816).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2020.106688.

References

[1] Dementia, World Health Organization, www.who.int/news-room/fact-sheets/detail/dementia, (Accessed on 01.08.2020).

[2] Alzheimer's Disease International, Dementia Statistics, www.alz.co.uk/research/statistics, (Accessed on 01.08.2020).

[3] R. Zhang, G. Simon, F. Yu, Advancing alzheimer's research: A review of big data promises, Int. J. Med. Inform. 106 (2017) 48–56, http://dx.doi.org/10.1016/j.ijmedinf.2017.07.002.

[4] 2016 Alzheimer's disease facts and figures, Alzheimer's Dementia 12 (4) (2016) 459–509, http://dx.doi.org/10.1016/j.jalz.2016.03.001.

[5] S. Iddi, D. Li, P.S. Aisen, M.S. Rafii, W.K. Thompson, M.C. Donohue, Predicting the course of Alzheime's progression, Brain Inform. 6 (2019) http://dx.doi.org/10.1186/s40708-019-0099-0.

[6] S. Lu, Y. Xia, W. Cai, M. Fulham, D.D. Feng, Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging, Comput. Med. Imaging Graph. 60 (2017) 35–41, http://dx.doi.org/10.1016/j.compmedimag.2017.01.001, Computational Methods for Molecular Imaging.

[7] K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, E. Katz, T. Nicholas, B. Billing, R. Anziano, T. Fullerton, Disease progression model for cognitive deterioration from alzheimer's disease neuroimaging initiative database, Alzheimer's dementia j. Alzheimer's Assoc. 7 (2) (2011) 151—160, http://dx.doi.org/10.1016/j.jalz.2010.03.018.

[8] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Modeling disease progression via multi-task learning, NeuroImage 78 (2013) 233–248, http://dx.doi.org/10.1016/j.neuroimage.2013.03.073.

[9] F. Liu, L. Zhou, C. Shen, J. Yin, Multiple kernel learning in the primal for multimodal Alzheimer's disease classification, IEEE J. Biomed. Health Inf. 18 (3) (2014) 984–990, http://dx.doi.org/10.1109/JBHI.2013.2285378.

[10] S. Duchesne, A. Caroli, C. Geroldi, D.L. Collins, G.B. Frisoni, Relating one-year cognitive change in mild cognitive impairment to baseline MRI features, NeuroImage 47 (4) (2009) 1363–1370, http://dx.doi.org/10.1016/j.neuroimage.2009.04.023.

[11] X. Wang, J. Qi, Y. Yang, P. Yang, A survey of disease progression modeling techniques for alzheimer's diseases, in: 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), 1, 2019, pp. 1237–1242, http://dx.doi.org/10.1109/INDIN41052.2019.8972091.

[12] A. Alberdi, A. Aztiria, A. Basarab, On the early diagnosis of Alzheimer's disease from multimodal signals: A survey, Artif. Intell. Med. 71 (2016) 1–29, http://dx.doi.org/10.1016/j.artmed.2016.06.003.

[13] G.M. Juan, G. Sanroma-Guell, G. Piella, A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease, Comput. Methods Programs Biomed. 189 (2020) 105348, http://dx.doi.org/10.1016/j.cmpb.2020.105348.

[14] S. Tabarestani, M. Aghili, M. Eslami, M. Cabrerizo, A. Barreto, N. Rishe, R.E. Curiel, D. Loewenstein, R. Duara, M. Adjouadi, A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study, NeuroImage 206 (2020) 116317, http://dx.doi.org/10.1016/j.neuroimage.2019.116317.

[15] D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, NeuroImage 59 (2) (2012) 895–907, http://dx.doi.org/10.1016/j.neuroimage.2011.09.069.

[16] Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables, NeuroImage 50 (4) (2010) 1519–1535, http://dx.doi.org/10.1016/j.neuroimage.2009.12.092.

[17] X. Ding, M. Bucholc, H. Wang, D.H. Glass, H. Wang, D.H. Clarke, A.J. Bjourson, L.R.C. Dowey, M. O'Kane, G. Prasad, L. Maguire, K. Wong-Lin, A hybrid computational approach for efficient Alzheimer's disease classification based on heterogeneous data, Sci. Rep. 8 (1) (2018) 9774, http://dx.doi.org/10.1038/s41598-018-27997-8.

[18] S. Qiu, G.H. Chang, M. Panagia, D.M. Gopal, R. Au, V.B. Kolachalama, Fusion of deep learning models of MRI scans, mini–mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment, in: Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 2018, pp. 737–749, http://dx.doi.org/10.1016/j.dadm.2018.08.013.

[19] L. Liu, S. Zhao, H. Chen, A. Wang, A new machine learning method for identifying Alzheimer's disease, Simul. Model. Pract. Theory 99 (2020) 102023, http://dx.doi.org/10.1016/j.simpat.2019.102023.

[20] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database, NeuroImage 56 (2) (2011) 766–781, http://dx.doi.org/10.1016/j.neuroimage.2010.06.013, Multivariate Decoding and Brain Reading.

[21] T. Wang, Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks, Sci. Rep. 8 (2018) http://dx.doi.org/10.1038/s41598-018-27337-w.

[22] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, NeuroImage 104 (2015) 398–412, http://dx.doi.org/10.1016/j.neuroimage.2014.10.002.

[23] P.S. Pillai, T.-Y. Leong, Fusing heterogeneous data for Alzheimer's disease classification, Stud. Health Technol. Inform. 216 (2015) 731–735, http://dx.doi.org/10.3233/978-1-61499-564-7-731.

[24] M. Ewers, C. Walsh, J.Q. Trojanowski, L.M. Shaw, R.C. Petersen, C.R. Jack, H.H. Feldman, A.L. Bokde, G.E. Alexander, P. Scheltens, B. Vellas, B. Dubois, M. Weiner, H. Hampel, Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance, Neurobiol. Aging 33 (7) (2012) 1203 – 1214.e2, http://dx.doi.org/10.1016/j.neurobiolaging.2010.10.019.

[25] K. Li, R. O'Brien, M. Lutz, S. Luo, T.A.D.N. Initiative, A prognostic model of Alzheimer's disease relying on multiple longitudinal measures and time-to-event data, Alzheimer's Dementia 14 (5) (2018) 644–651, http://dx.doi.org/10.1016/j.jalz.2017.11.004.

[26] P. Forouzannezhad, A. Abbaspour, C. Li, C. Fang, U. Williams, M. Cabrerizo, A. Barreto, J. Andrian, N. Rishe, R.E. Curiel, D. Loewenstein, R. Duara, M. Adjouadi, A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging, J. Neurosci. Methods 333 (2020) 108544, http://dx.doi.org/10.1016/j.jneumeth.2019.108544.

[27] G. Lee, K. Nho, B. Kang, K.-A. Sohn, D. Kim, Predicting Alzheimer's disease progression using multi-modal deep learning approach, Sci. Rep. 9 (2019) 1952, http://dx.doi.org/10.1038/s41598-018-37769-z.

[28] W. Liu, B. Zhang, Z. Zhang, X.-H. Zhou, Joint modeling of transitional patterns of Alzheimer's disease, PLOS ONE 8 (9) (2013) 1–11, http://dx.doi.org/10.1371/journal.pone.0075487.

[29] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, D. Shen, Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, Neurobiol. Aging 46 (2016) 180–191, http://dx.doi.org/10.1016/j.neurobiolaging.2016.07.005.

[30] M. Liu, J. Zhang, E. Adeli, D. Shen, Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis, IEEE Trans. Biomed. Eng. 66 (5) (2019) 1195–1206, http://dx.doi.org/10.1109/TBME.2018.2869989.

[31] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, X. Li, Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease, IEEE Trans. Neural Netw. Learn. Syst. 28 (7) (2017) 1508–1519, http://dx.doi.org/10.1109/TNNLS.2016.2520964.

[32] Q. Liao, Y. Ding, Z.L. Jiang, X. Wang, C. Zhang, Q. Zhang, Multi-task deep convolutional neural network for cancer diagnosis, Neurocomputing 348 (2019) 66–73, http://dx.doi.org/10.1016/j.neucom.2018.06.084, Advances in Data Representation and Learning for Pattern Analysis.

[33] H. Harutyunyan, H. Khachatrian, D.C. Kale, A. Galstyan, Multitask learning and benchmarking with clinical time series data, in: Scientific Data, (2019), 6, 2019, http://dx.doi.org/10.1038/s41597-019-0103-9.

[34] Y. Zhang, Q. Yang, A survey on multi-task learning, (2017), 2017, CoRR abs/1707.08114, arXiv:1707.08114.

[35] Y. Cho, J.-K. Seong, Y. Jeong, S.Y. Shin, Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data, NeuroImage 59 (3) (2012) 2217–2230, http://dx.doi.org/10.1016/j.neuroimage.2011.09.085.

[36] B.M. Jedynak, A. Lang, B. Liu, E. Katz, Y. Zhang, B.T. Wyman, D. Raunig, C.P. Jedynak, B. Caffo, J.L. Prince, A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort, NeuroImage 63 (3) (2012) 1478–1486, http://dx.doi.org/10.1016/j.neuroimage.2012.07.059.

[37] W.-Y.W. Yau, D.L. Tudorascu, E.M. McDade, S. Ikonomovic, J.A. James, D. Minhas, W. Mowrey, L.K. Sheu, B.E. Snitz, L. Weissfeld, P.J. Gianaros, H.J. Aizenstein, J.C. Price, C.A. Mathis, O.L. Lopez, W.E. Klunk, Longitudinal assessment of neuroimaging and clinical markers in autosomal dominant Alzheimer's disease: a prospective cohort study, Lancet Neurol. 14 (8) (2015) 804–813, http://dx.doi.org/10.1016/S1474-4422(15)00135-0.

[38] R. Guerrero, A. Schmidt-Richberg, C. Ledig, T. Tong, R. Wolz, D. Rueckert, Instantiated mixed effects modeling of Alzheimer's disease markers, NeuroImage 142 (2016) 113–125, http://dx.doi.org/10.1016/j.neuroimage.2016.06.049.

[39] M. Mehdipour Ghazi, M. Nielsen, A. Pai, M.J. Cardoso, M. Modat, S. Ourselin, L.S. rensen, Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling, Med. Image Anal. 53 (2019) 39–46, http://dx.doi.org/10.1016/j.media.2019.01.004.

[40] H. Suresh, N. Hunt, A. Johnson, L.A. Celi, P. Szolovits, M. Ghassemi, Clinical Intervention Prediction and Understanding with Deep Neural Networks, in: Proceedings of the 2nd Machine Learning for Healthcare Conference, vol. 68, 2017, pp. 322–337.

[41] C. Tian, J. Ma, C. Zhang, P. Zhan, A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network, Energies 11 (2018) 3493, http://dx.doi.org/10.3390/en11123493.

[42] S. Spasov, L. Passamonti, A. Duggento, P. Lió, N. Toschi, A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease, NeuroImage 189 (2019) 276–287, http://dx.doi.org/10.1016/j.neuroimage.2019.01.031.

[43] S. Tabarestani, M. Aghili, M. Shojaie, C. Freytes, M. Cabrerizo, A. Barreto, N. Rishe, R.E. Curiel, D. Loewenstein, R. Duara, M. Adjouadi, Longitudinal Prediction Modeling of Alzheimer Disease using Recurrent Neural Networks, in: 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2019, pp. 1–4, doi:10.1109/BHI.2019.8834556.

[44] D.A. Llano, G. Laforet, V. Devanarayan, Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to alzheimer disease, Alzheimer Dis. Assoc. Disord. 25 (1) (2011) 73—84, http://dx.doi.org/10.1097/wad.0b013e3181f5b8d8.

[45] G. Frisoni, N. Fox, C. Jack, P. Scheltens, P. Thompson, The clinical use of structural MRI in Alzheimer's disease, Nature Rev. Neurol. 6 (2010) 67–77, http://dx.doi.org/10.1038/nrneurol.2009.215.

[46] B. Cheng, M. Liu, D. Zhang, B.C. Munsell, D. Shen, Domain transfer learning for MCI conversion prediction, IEEE Trans. Biomed. Eng. 62 (7) (2015) 1805–1817, http://dx.doi.org/10.1109/TBME.2015.2404809.

[47] C.-Y. Wee, P.-T. Yap, D. Shen, Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns, Hum. brain mapp. 34 (2013) http://dx.doi.org/10.1002/hbm.22156.

[48] P.J. Moore, T.J. Lyons, J. Gallacher, for the random forest prediction of Alzheimer's disease using pairwise selection from time series data, PLOS ONE 14 (2) (2019) 1–14, http://dx.doi.org/10.1371/journal.pone.0211558.

[49] T. Yagi, M. Kanekiyo, J. Ito, R. Ihara, K. Suzuki, A. Iwata, T. Iwatsubo, K. Aoshima, Identification of prognostic factors to predict cognitive decline of patients with early alzheimer's disease in the Japanese alzheimer's disease neuroimaging initiative study, in: Alzheimer's & Dementia: Translational Research & Clinical Interventions, 5, 2019, pp. 364–373, http://dx.doi.org/10.1016/j.trci.2019.06.004.

[50] H. Choi, K.H. Jin, Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging, Behav. Brain Res. 344 (2018) 103–109, http://dx.doi.org/10.1016/j.bbr.2018.02.017.

[51] N. Amoroso, D. Diacono, A. Fanizzi, M. La Rocca, A. Monaco, A. Lombardi, C. Guaragnella, R. Bellotti, S. Tangaro, Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge, J. Neurosci. Methods 302 (2018) 3–9, http://dx.doi.org/10.1016/j.jneumeth.2017.12.011, A machine learning neuroimaging challenge for automated diagnosis of Alzheimer's disease.

[52] L. Sun, B. Zou, S. Fu, J. Chen, F. Wang, Speech emotion recognition based on DNN-decision tree SVM model, Speech Commun. 115 (2019) 29–37, http://dx.doi.org/10.1016/j.specom.2019.10.004.

[53] S. Notley, M. Magdon-Ismail, Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and K-nearest neighbor classifiers, 2018, CoRR abs/1805.02294, arXiv:1805.02294.

[54] D. Maji, A. Santara, S. Ghosh, D. Sheet, P. Mitra, Deep neural network and random forest hybrid architecture for learning to detect retinal vessels in fundus images, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 3029–3032, http://dx.doi.org/10.1109/EMBC.2015.7319030.

[55] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, Discriminative self-representation sparse regression for neuroimaging-based alzheimer's disease diagnosis, Brain imaging behav. 13 (1) (2019) 27–40, http://dx.doi.org/10.1007/s11682-017-9731-x.

[56] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment, Front. Neurosci. 12 (2018) 777, http://dx.doi.org/10.3389/fnins.2018.00777.

[57] X. Hong, R. Lin, C. Yang, N. Zeng, C. Cai, J. Gou, J. Yang, Predicting Alzheimer's disease using LSTM, IEEE Access 7 (2019) 80893–80901, http://dx.doi.org/10.1109/ACCESS.2019.2919385.

[58] M. Bucholc, X. Ding, H. Wang, D.H. Glass, H. Wang, G. Prasad, L.P. Maguire, A.J. Bjourson, P.L. McClean, S. Todd, D.P. Finn, K. Wong-Lin, A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual, Expert Syst. Appl. 130 (2019) 157–171, http://dx.doi.org/10.1016/j.eswa.2019.04.022.

[59] N. Bhagwat, J. Pipitone, A. Voineskos, M. Chakravarty, An artificial neural network model for clinical score prediction in alzheimer disease using structural neuroimaging measures, J. Psychiatry. Neurosci. 44 (2019) 246–260, http://dx.doi.org/10.1503/jpn.180016.

[60] J.J. Gomar, C. Conejero-Goldberg, P. Davies, T.E. Goldberg, Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data, Alzheimer's Dementia 10 (6) (2014) 704–712, http://dx.doi.org/10.1016/j.jalz.2013.11.009.

[61] F. Jiménez, G. Sánchez, J.M. Juárez, Multi-objective evolutionary algorithms for fuzzy classification in survival prediction, Artif. Intell. Med. 60 (3) (2014) 197–219, http://dx.doi.org/10.1016/j.artmed.2013.12.006.

[62] R. Babapour, N. Schoonenboom, B. Tijms, P. Scheltens, P. Visser, W. Flier, C. Teunissen, Decision tree supports the interpretation of CSF biomarkers in alzheimer's disease, in: Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 11, 2018, http://dx.doi.org/10.1016/j.dadm.2018.10.004.

[63] D. Das, J. Ito, T. Kadowaki, K. Tsuda, An interpretable machine learning model for diagnosis of Alzheimer's disease, vol. 7, PeerJ, 2019, e6543, http://dx.doi.org/10.7717/peerj.6543.

[64] D. Chitradevi, S. Prabha, Analysis of brain sub regions using optimization techniques and deep learning method in Alzheimer disease, Appl. Soft Comput. 86 (2020) 105857, http://dx.doi.org/10.1016/j.asoc.2019.105857.

[65] G. Uysal, M. Ozturk, Hippocampal atrophy based Alzheimer's disease diagnosis via machine learning methods, J. Neurosci. Methods 337 (2020) 108669, http://dx.doi.org/10.1016/j.jneumeth.2020.108669.

[66] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681, http://dx.doi.org/10.1109/78.650093.

[67] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[68] M. Elkano, M. Galar, J. Sanz, H. Bustince, Fuzzy rule-based classification systems for multi-class problems using binary decomposition strategies: On the influence of n-dimensional overlap functions in the fuzzy reasoning method, Inform. Sci. 332 (2016) 94–114, http://dx.doi.org/10.1016/j.ins.2015.11.006.

[69] D.P. Pancho, J.M. Alonso, L. Magdalena, Enhancing fingrams to deal with precise fuzzy systems, Fuzzy Sets and Systems 297 (2016) 1–25, http://dx.doi.org/10.1016/j.fss.2015.05.019, Themed Section: Fuzzy Systems.

[70] F. Li, L. Tran, K. Thung, S. Ji, D. Shen, J. Li, A robust deep model for improved classification of AD/MCI patients, IEEE J. Biomed. Health Inf. 19 (5) (2015) 1610–1616, http://dx.doi.org/10.1109/JBHI.2015.2429556.

[71] H. Liedes, J. Mattila, M. Gils, J. Koikkalainen, H. Soininen, J. Lötjönen, Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort, J. Alzheimer's dis. 39 (2013) http://dx.doi.org/10.3233/JAD-130359.

[72] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (2016) http://dx.doi.org/10.1038/s41598-018-24271-9.

[73] Y. Zhang, P.J. Thorburn, W. Xiang, P. Fitch, SSIM—A Deep learning approach for recovering missing time series sensor data, IEEE Internet Things J. 6 (4) (2019) 6618–6628, http://dx.doi.org/10.1109/JIOT.2019.2909038.

[74] R. Cui, M. Liu, RNN-Based longitudinal analysis for diagnosis of alzheimer's disease, Comput. Med. Imaging Graph. 73 (2019) 1–10, http://dx.doi.org/10.1016/j.compmedimag.2019.01.005.

[75] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321–357.

[76] A. Elisseeff, M. Pontil, Leave-one-out error and stability of learning algorithms with applications stability of randomized learning algorithms source, Int. J. Syst. Sci. IJSySc 6 (2002).

[77] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: Y. Bengio, Y. LeCun ((Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings,.

[78] X. Zhao, Y. Wu, D.L. Lee, W. Cui, Iforest: Interpreting random forests via visual analytics, IEEE Trans. Vis. Comput. Graphics 25 (1) (2019) 407–416, http://dx.doi.org/10.1109/TVCG.2018.2864475.

[79] H. Deng, Interpreting tree ensembles with intrees, Int. J. Data Sci. Anal. 7 (2019) 277–287, http://dx.doi.org/10.1007/s41060-018-0144-8.

[80] Y. Ming, H. Qu, E. Bertini, Rulematrix: Visualizing and understanding classifiers with rules, IEEE Trans. Vis. Comput. Graphics 25 (1) (2019) 342–352, http://dx.doi.org/10.1109/TVCG.2018.2864812.

[81] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '15, 2015, pp. 1721–1730, http://dx.doi.org/10.1145/2783258.2788613.

[82] K. Ritter, J. Schumacher, M. Weygandt, R. Buchert, C. Allefeld, J.-D. Haynes, Multimodal prediction of conversion to alzheimer's disease based on incomplete biomarkers, in: Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1, (2) 2015, pp. 206–215, http://dx.doi.org/10.1016/j.dadm.2015.01.006.

[83] S. El-Sappagh, T. Abuhmed, S. Riazul Islam, K.S. Kwak, Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data, Neurocomputing 412 (2020) 197–215, http://dx.doi.org/10.1016/j.neucom.2020.05.087.

[84] K. Oh, Y.-C. Chung, K. Kim, W.-S. Kim, I.-S. Oh, Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning, Sci. Rep. 9 (2019) http://dx.doi.org/10.1038/s41598-019-54548-6.

[85] I. Beheshti, H. Demirel, H. Matsuda, Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm, Comput. Biol. Med. 83 (2017) 109–119, http://dx.doi.org/10.1016/j.compbiomed.2017.02.011.

[86] C. Fang, C. Li, P. Forouzannezhad, M. Cabrerizo, R. Curiel, D. Loewenstein, R. Duara, M. Adjouadi, Gaussian Discriminative component analysis for early detection of Alzheimer's disease: A supervised dimensionality reduction algorithm, J. Neurosci. Methods 344 (2020) 108856, http://dx.doi.org/10.1016/j.jneumeth.2020.108856.

[87] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, NeuroImage 208 (2020) 116459, http://dx.doi.org/10.1016/j.neuroimage.2019.116459.

[88] S. Dimitriadis, D. Liparas, M. Tsolaki, Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate healthy elderly, early/late MCI and Alzheimer's disease patients: from Alzheimer's disease neuroimaging initiative (ADNI) database, J. Neurosci. Methods 302 (2017) http://dx.doi.org/10.1016/j.jneumeth.2017.12.010.

[89] C. Platero, M. Tobar, Longitudinal survival analysis and two-group comparison for predicting the progression of mild cognitive impairment to Alzheimer's disease, J. Neurosci. Methods 341 (2020) 108698, http://dx.doi.org/10.1016/j.jneumeth.2020.108698.

[90] A. Abrol, M. Bhattarai, A. Fedorov, Y. Du, S. Plis, V. Calhoun, Deep residual learning for neuroimaging: An application to predict progression to Alzheimer's disease, J. Neurosci. Methods 339 (2020) 108701, http://dx.doi.org/10.1016/j.jneumeth.2020.108701.

[91] D. Lu, K. Popuri, G.W. Ding, R. Balachandar, M.F. Beg, Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease, Med. Image Anal. 46 (2018) 26–34, http://dx.doi.org/10.1016/j.media.2018.02.002.

[92] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, NeuroImage Clin. 21 (2019) 101645, http://dx.doi.org/10.1016/j.nicl.2018.101645.

[93] A. Zandifar, V.S. Fonov, S. Ducharme, S. Belleville, D.L. Collins, MRI And cognitive scores complement each other to accurately predict Alzheimer's dementia 2 to 7 years before clinical onset, NeuroImage Clin. 25 (2020) 102121, http://dx.doi.org/10.1016/j.nicl.2019.102121.

[94] K.R. Kruthika, H.D. Rajeswari, Maheshappa, Multistage classifier-based approach for Alzheimer's disease prediction and retrieval, Inform. Med. Unlocked 14 (2019) 34–42, http://dx.doi.org/10.1016/j.imu.2018.12.003.

[95] K. Huang, Y. Lin, L. Yang, Y. Wang, S. Cai, L. Pang, X. wu, L. Huang, A multipredictor model to predict the conversion of mild cognitive impairment to Alzheimer's disease by using a predictive nomogram, Neuropsychopharmacology 45 (2019) http://dx.doi.org/10.1038/s41386-019-0551-0.

[96] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, X. Song, Early detection of Alzheimer's disease using magnetic resonance imaging: A novel approach combining convolutional neural networks and ensemble learning, Front. Neurosci. 14 (2020) 259, http://dx.doi.org/10.3389/fnins.2020.00259.

[97] S. El-Sappagh, H. Saleh, R. Sahal, T. Abuhmed, S.R. Islam, F. Ali, E. Amer, Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data, Future Gener. Comput. Syst. 115 (2021) 680–699, http://dx.doi.org/10.1016/j.future.2020.10.005.