Neurocomputing 512 (2022) 203–224

Contents lists available at ScienceDirect

Neurocomputing

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

Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers

Shaker El-Sappagh [a,b,c,1], Farman Ali [d,1], Tamer Abuhmed [c,*], Jaiteg Singh [e], Jose M. Alonso [f,*]

[a] Faculty of Computer Science and Engineering, Galala University, 435611 Suez, Egypt
[b] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, 13518 Banha, Egypt
[c] College of Computing and Informatics, Sungkyunkwan University, South Korea
[d] Department of Software, Sejong University, Seoul, South Korea
[e] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[f] Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 25 August 2021
Revised 29 June 2022
Accepted 4 September 2022
Available online 20 September 2022

Keywords:
Computational Intelligence
Data fusion
Ensemble classifiers
Stacking
Data analysis
Alzheimer disease progression detection

ABSTRACT

Predicting Alzheimer's disease (AD) progression is crucial for improving the management of this chronic disease. Usually, data from AD patients are multimodal and time series in nature. This study proposes a novel ensemble learning framework for AD progression incorporating heterogeneous base learners into an integrated model using the stacking technique. This framework is used to build a 4-class ensemble classifier, which predicts AD progression 2.5 years in the future based on the multimodal time-series data. Statistical measures have been extracted from the longitudinal data to be used by the conventional machine learning models. The examined ensemble members include k-nearest neighbor, extreme gradient boosting, support vector machine, random forest, decision tree, and multilayer perceptron. We utilize three time-series modalities and one static non-time series modality of 1371 subjects from the Alzheimer's disease neuroimaging initiative (ADNI) to validate our model. Several homogeneous and heterogeneous combinations of ensemble members were implemented, and their performance compared. The balance between accuracy and diversity when selecting ensemble members was investigated. We found that both accuracy and diversity are equally critical metrics to obtain an optimal ensemble model. Furthermore, our testing showed that the proposed model achieves outstanding progression prediction performance. The proposed model achieved a high performance without using neuroimaging data, which means that the model could be implemented in low-cost healthcare environments. The proposed model has achieved superior results compared with the state-of-the-art techniques in Alzheimer's and ensemble classifiers domains. The proposed framework can be used to implement efficient information fusion ensembles for other medical and non-medical problems.

1. Introduction

Alzheimer's disease (AD) is an irreversible and incurable neurodegenerative disease. It is the most common form of dementia in the elderly [1]. AD has debilitating effects on patients, their families, society and causes immeasurable grief to caregivers. Mild cognitive impairment (MCI) is an intermediate stage between cognitively normal and dementia. Around 10 % to 20 % of MCI patients progress to AD within one year [2,3], the remaining non-converting subjects either develop other forms of dementia or remain stable. Early identification of patients who will convert to AD before the appearance of advanced symptoms is crucial for proper treatment and remains a significant medical challenge. Progression detection potentially enables early AD diagnosis and allows the medication to be administered to slow down the disease's progression. However, even discriminating stable MCI (sMCI) from progressive MCI (pMCI) patients is challenging because these patients have similar properties. The combination of these two challenging problems yields a much more complicated 4-class (cognitive normal [CN] vs sMCI vs pMCI vs AD) classification problem for any AD progression detection approach.

Recently, many studies and competitions have been proposed various approaches to solve this problem using a combination of markers and biomarkers, e.g., positron emission tomography (PET) or magnetic resonance imaging (MRI) data as inputs to

* Corresponding authors.
E-mail addresses: tamer@skku.edu (T. Abuhmed), josemaria.alonso.moral@usc.es (J.M. Alonso).
[1] These authors contributed equally to this work and co-first authors.

https://doi.org/10.1016/j.neucom.2022.09.009
0925-2312/

machine learning (ML) algorithms [1,4–6]. These studies were based on a limited set of features from a few modalities, and many relied especially heavily on neuroimaging data. In this context, Bron et al. [7] proposed the CADDementia grand challenge[2] based on an MRI dataset of 384 patients. The challenge involves classification to one of three classes (CN vs MCI vs AD), i.e., there is no differentiation between sMCI and pMCI. By achieving an accuracy of 63.0 %, Sorensen et al. [8] won this challenge using a linear discriminant analysis (LDA) classifier fed with MRI volumetric, cortical thickness, hippocampal texture, and hippocampal shape features. Recently, Kaggle[3] proposed another competition for AD prediction based on a set of MRI modality features plus three demographics. This challenge was presented as a 4-class classification problem (CN vs pMCI vs sMCI vs AD). Again, the number of cases utilized was limited, as the training and test sets consisted of only 240 and 160 subjects, respectively. As expected, the performance of the resulting models was worse than in the CADDementia challenge because increasing the number of classes makes the classification problem harder. For example, Yao et al. [6] proposed a hierarchical ensemble classifier that converts this 4-class classification problem into five binary classification problems, achieving 54.38 % accuracy. Dimitriadis et al. [9] built five ensemble classifiers based on the random forest classification technique to achieve an AD diagnosis as the result of a 4-class classification problem. They achieved an accuracy of 61.9 % and asserted that this performance was the best recorded in the literature to that date.

Although there have been many AD classification and progression studies that have used a wide variety of data types, including MRI [10–12], PET [13], cerebrospinal fluid (CSF) biomarkers [14,15], or fusion of neuro-images [16], there is still a significant gap between research results and being able to apply these systems in daily clinical practice [17,18]. The deployment of ML models for real-time AD diagnoses is still limited because the results of previous studies have been inconclusive. This is mainly due to the complexity of the problem, the inefficient utilization of available data, and the ineffective design of the ML models [19]. There has been an increasing interest in multivariate approaches [20], where information fusion plays a crucial role in transforming heterogeneous patient data into high-quality fused data [21,22]. By taking this approach, we can provide a complete picture of patient status. Most existing studies mainly depend on extracted features from MRI neuroimaging, but these have failed to accurately predict MCI-to-AD progression [20,21]. Donnelly-Kehoe et al. [23] concluded that the maximum accuracy achieved by using MRI features does not reach the standard of using the mini-mental state examination (MMSE) by itself. In addition, because AD progression classification is medically very complex, depending only on MRI is not acceptable to physicians. In healthcare environments, physicians usually integrate many features from different modalities to make their decisions. Hence, it is essential we explore fusing multiple modalities such as cognitive score markers (CSs), MRI, neuropsychological battery markers (NSB), and CSF biomarkers [24–26] to improve the accuracy of classification models. However, the effect of multimodal fusion needs further study.

Since AD is a chronic disease, a patient's condition keeps evolving over time, and this is because patients are experiencing gradual structural and functional changes to their brain. Most previous studies have predicted the progression of MCI based on baseline visit data only [26]. Thus, they fail to fully represent or consider time-variant medical records that could facilitate comprehensive predictive ML models. Traditional time series methods, such as hidden Markov chains, have been used to study the AD progression detection problem [27–30]. As the maximum number of visits for a patient in our dataset is four, traditional time series methods are not suitable for capturing disease progression patterns [31]. Deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), are also commonly used for time series data analysis, but they require a large amount of data for model training [22,21]. Robust hybrid deep learning models have already been successfully applied to AD progression detections [21]. Moreover, in the medical domain, it is not easy to introduce novel ML methods while physicians are asking for methods that are multi-modal with comprehensible recommendations [32]. Unfortunately, neural network-based ML methods act as non-understandable black boxes and behave poorly when considering small or medium-sized datasets outside of image processing. In this work, we avoid using the neural network-based ML methods and only consider the interpretable ML models such as the decision tree, support vector machines, and random forest.

The role of time series data to improve ML models in the AD progression domain deserves more attention. Authors have published several research works in AD domain [18,21,22,32,33]. In [21], we proposed two novel deep learning models for AD progression detection. These models were based on the fusion of multimodal time series data of four-time steps. One model is a hybrid model of Bidirectional LSTM and random forest, and the other model is based on multitask modeling to predict seven cognitive scores. These scores were used to predict AD progression after 2.5 years. In [22], we proposed a novel deep learning model called the stacked CNN-LSTM model. The technique was based on a multitask modeling approach to jointly predict multiple cognitive scores (regression tasks) and concurrently AD progression (multiclass classification task) within 7.5 years. The study mainly concentrated on studying the role of adding multiple time steps to enhance the performance of a deep learning model and the role of multitasking to build a stable and reliable deep learning model. We concentrated mainly on proposing a deep learning model that can learn complex and multimodal time series data of 15-time steps. In [32], we tried to provide a more understandable decision for AD progression detection by providing explainability capabilities to our model. We used the random forest to build a two-layer model, i.e., the first layer was to early diagnose AD patients, and the second layer was to detect possible MCI-to-AD progression. In [18], we tried to build a cost-effective machine learning model for AD progression detection. The model has been based on a set of cheap features that can be easily collected from patient's electronic health record at a low cost. As far as we know, the application of ML algorithms using multimodal and time-series data has not yet achieved good results [34]. Ritter et al. [1] combined ten modalities, including MRI, PET, and NSB, to predict AD progression within three years. They achieved an accuracy of 73 % using support vector machines. The generalization error of ML models can be divided into three parts: $error = bias^2 + variance + noise$. An ensemble system, also known as multiple classifier system, combines a pool of intelligent classifiers seeking to exploit the strengths of each classifier in such a way as to reduce the generalization error you may get from any single model [35]. For example, bagging methods use undersampling to minimize the variance of the base models. Boosting methods reduce bias using a weighting strategy. Subspace methods divide the feature vector space into several subspaces, this improves the diversity of the trained models and reduces noise. Based on this data fusion from multiple sources, an ensemble classifier can become more accurate and outperform the best base classifiers [36,37]. Ensemble models have been used for years in different applications, including in AD progression

[2] *CADDementia grand challenge: https://caddementia.grand-challenge.org/.*
[3] *KAGGLE competition on MCI prediction: https://www.kaggle.com/c/mci-prediction/.*

prediction [6,38,39]. In the TADPOLE grand challenge,[4] Moore et al. [40] applied the random forest technique to predict AD achieving an AUC of 0.82 and a 3-class classification accuracy of 0.73. Most AD studies have been based on small datasets using baseline MRI features. Farhan et al. [38] used only five features from 85 MRI scans to build an ensemble of classifiers. Unfortunately, these models are not trusted in the medical domain due to a lack of understanding of how the conclusions were drawn [1]. Although ensemble models are expected to achieve better performance than base models [35], this optimal performance usually implies an increase in the complexity of the model that makes it hard for physicians to interpret. Moreover, optimal performance is not guaranteed simply using an ensemble model, and an appropriate ensemble architecture must be carefully designed. For example, (1) the most informative and discriminative features must be selected; (2) suitable data and decision fusion strategies must be chosen; and (3) the most accurate and highly diverse, least correlated, as well as mutually complementary base learners must be implemented. However, it is critical to select the best set of base classifiers and then integrating them in an optimal way for creating an accurate and stable ensemble. Yu and Zhao [41] proposed the Bayesian network-based probabilistic ensemble learning strategy to automatically and effectively embed the suitable base classifiers in a Bayesian network in a probabilistic manner. The study highlighted the importance of selecting suitable base classifiers, and they used the Bayesian network to take this decision. It is worth noting that current AD ensemble-based studies tend to utilize a limited amounts of training data, feature sets, and numbers of modalities while ignoring time series data completely [40,42].

To overcome these limitations that are found in current AD detection approaches, this work proposes a methodology to find the best ensemble classifier by looking at six types of classification models, namely the decision tree (DT), support vector machines (SVM), random forest (RF), k-NN (KNN), multilayer perceptron (MLP), and extreme gradient boosting (XGB) models [40,43–45]. This study considers ensemble models that are based on the fusion of heterogeneous yet complementary time-series data from different AD modalities. Unlike existing literature, we used AD patient's multimodal medical information from various time points to predict the disease's progression at a future time point. The study optimized this set of well-known machine learning techniques. The proposed models were optimized using a collection of cost-effective time-series features including patient's comorbidities, cognitive scores, medication history, and demographics. The current study implemented a different machine leaning technique (ensemble classifiers), explored different feature set to optimize ensemble models, perform critical statistical analysis for feature engineering, and prove the theory of diversity vs accuracy to build a highly performing ensemble classifier for AD progression detection. We built information fusion enabled ensemble classifiers based on a patient's first four visits (baseline [BL], month-6 [M06], month-12 [M12], and month-18 [M18]) to predict the patient's AD state 2.5 years the final visit (i.e., at month-48). It is challenging to select the best set of features from a patient's highly heterogeneous multimodal data. These features should form a representative, informative, and discriminative list that best describes the temporal patterns at play in the time-series data. In addition, we must integrate these time series features with other baseline static features such as age, number of years in education, etc. It should also be noted that selecting a subset of base classifiers based only on their individual performance does not guarantee optimum performance for the resulting ensemble. Diversity of base learners is crucial for implementing an accurate ensemble [36]; therefore, we studied the accuracy/diversity trade-off and its role in selecting candidate base learners from a pool of classifiers. A single measure of diversity is inadequate [46], so we implemented three pair-wise measures (Q-statistic, correlation, and disagreement) to test the diversity among fused models. We also implemented a list of ensemble frameworks to study the effect of early and late data fusion schemes on the ensembles' performance. Then, we select the best ensemble model and optimize its parameters. *To the best of our knowledge, this is the first study that builds an ensemble model for multiclass AD prediction detection based on multimodal time series data.* The main contributions of this study are as follows:

- An ensemble learning framework based on extensive multimodal time series data is proposed for AD progression detection. The proposed framework integrates early data fusion and late decision fusion features to create more stable and accurate ensemble classifiers.
- Statistical analysis is conducted for the feature extraction process using four medically critical modalities (i.e., CSs + MRI + NSB + Static). These modalities are time-series data with 4-time timesteps (i.e., BL + M06 + M12 + M18).
- Information gain and recursive feature elimination strategies are applied to select the most discriminative features from time series and static modalities.
- An actual patient data from 1371 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI)[5] is used to optimize the proposed AD progression prediction ensemble model. The model predicts MCI-to-AD progression 2.5 years after the last visit. We explore the role of different AD modality combinations on AD disease progression detection.
- Extensive experiments are performed to evaluate the effectiveness of a set of proposed ensemble models based on complete patient profiles. We also studied the behavior of different ensemble techniques, including voting and stacking. In addition, we compared the behavior of these ensembles with base models such as RF, XGB, SVM, MLP, DT, and KNN.
- We compare the homogeneous and heterogeneous designs of ensemble models and measure the effects of that on the ensemble's performance.
- We explore the effect of changing the diversity and performance of base classifiers on the overall performance of the ensemble. In other words, we evaluate the tradeoff between diversity and accuracy of based models on the overall performance of ensembles.
- To implement the optimum ensemble model, we consider both the medical relationship among selected features from each modality and the accuracy/diversity trade-off of the base classifiers. Accordingly, we chose a discriminating list of features, the best fusion scheme, and the most accurate and diverse subset of base classifiers.
- The hyperparameters of the selected stacking ensemble framework are optimized using a random search method and real-world data from the ADNI. The experimental results confirm the effectiveness of this stacking model for AD progression detection.

The rest of the paper is organized as follows. Section 2 discusses our methodology and introduces the proposed framework. Section 3 discusses the experimental results. Section 4 concludes the paper and provides readers with hints on future directions to extend the proposed models.

[4] *TADPOLE grand challenge:* https://tadpole.grand-challenge.org/.

[5] Alzheimer's Disease Neuroimaging Initiative (ADNI): https://adni.loni.usc.edu/.

2. Methods

2.1. ADNI study

The dataset used in this study is the ADNI[6] database. The database was accessed on March 18, 2019. We selected 1371 subjects (54.5 % male) from ADNI 1, ADNI GO, and ADNI 2 based on the available data. Subjects are categorized into four groups based on the individual's clinical diagnosis at the baseline and at future-time points, this is illustrated in Fig. 1: (1) CN: 419 subjects diagnosed to be CN at baseline and remained CN over all future time steps; (2) sMCI: 473 subjects diagnosed to have MCI at all time-points; (3) pMCI: 140 subjects that had MCI at baseline + M06 + M12 + M 18 visits but were predicted to progress to AD within 2.5 years from M18 (i.e., by the M48 visit); (4) AD: 339 subjects diagnosed as having AD in all visits. A list of patient IDs used can be found in Supplementary File 1. Any subject that showed improvement in his/her diagnoses during follow-ups was excluded from the study. In other words, those clinically diagnosed as having MCI but reverted to being CN, or those clinically diagnosed as having AD but reverted to having MCI or being CN are considered to have been misdiagnosed as AD is an irreversible form of dementia. Furthermore, cases that where direct conversion from CN to AD occurs are also discarded.

2.2. Selected raw features

Several types of biomarker and neuropsychological test entries are included in the ADNI data and have been individually validated for AD progression. In accordance with results reported in previous research studies in the specialized literature, the following features are considered as potential predictors for AD progression (we have organized these raw features into two main groups) [1,2,6,12,17,31,24,40].

(1) The first group of features are longitudinal and are considered as time series data. These data were collected at baseline and then regularly every-six months up to month 18 (i.e., four entries over 1.5 years, see Fig. 1). In the ADNI, time-series data are medically divided into the following three modalities of (1) Cognitive scores CSs (9 features); (2) neuropsychological battery NSB (51 features); and (3) MRI scans (312 features). The CS features include Alzheimer's disease assessment scales (ADAS 11 and ADAS 13), clinical dementia rating-sum of boxes (CDRSB), global CDR (CDGLOBAL), functional assessment questionnaire (FAQ), geriatric depression scale (GDTOTAL), MMSE, Montreal cognitive assessment (MoCA), and neuropsychiatric inventory score (NPISCORE). Note that we only use sum scores of different neuropsychological tests to cover the important aspects of those tests [1]. The NSB features include Rey's auditory verbal learning test (RAVLT) features, everyday cognition reports, etc. The imaging data used in our experiments are based on a preprocessed set of T1 weighted MRI features from the ADNI database. The data was preprocessed with the standard ADNI pipeline by a team from the University of California at San Francisco (UCSF), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer[7] image analysis suite according to the atlas generated in Desikan et al. [47]. The MRI features are associated with regional cortical thickness, regional volume, and surface area measures. Details of the analysis procedure are available online[8] and upon request from the authors. Full details about these features can be found at the ADNI website, the ADNI reference repository,[9] and Supplementary File 2. It is worth noting that PET imaging is also available but has some limitations such as (i) the number of scans is small because the patient is exposed to ionizing radiation, (ii) it has a lower spatial resolution in comparison with MRI, and (iii) it is extremely expensive. Therefore, to build a practical model, we ignore the PET data.

(2) The second group of features is static one-time baseline data collected only once (at the baseline visit). These data covers 64 features, including CSF biomarker (3 features of the amyloid-β peptide from 42 amino acids-$A\beta_{1-42}$ [ABETA], TAU, and phosphorylated TAU [PTAU]); genetic information (one feature from APOe4); family history questionnaire (5 features from family member dementia history); CSF local lab results (4 features from blood cell counts as well as protein and glucose results); symptoms (27 features from occurrences of nausea, crying, falls, etc.); sociodemographic (5 features from age, marriage status, body mass index [BMI], gender, and the number of years education data); and medical history (19 features from psychiatric, neurologic, cardiovascular, musculoskeletal, gastrointestinal, etc.).

ADNI subjects are 55 to 92 years old, fluent in English or Spanish, and had at least six years of education. At baseline, subjects met the following specific inclusion criteria described in Table 1. They met the diagnostic criteria for probable AD by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [48]. Supplementary File 1 has a complete overview of the data types used in our study. Detailed descriptions of the ADNI subjects, image acquisition protocols and procedures, as well as post-acquisition preprocessing procedures can be found at *https://www.adni-info.org* and upon request from the authors. Demographic and clinical information from the subjects is shown in Table 1.

2.3. Model design and development

Fig. 2 illustrates the pipeline associated with our model. As introduced previously, we have two types of data (i.e., time-series data, and static data). The whole pipeline is comprised of seven stages.

- *First*, the entire dataset is randomly split, with 80 % (1097 patients) going into the model development set (MDS) and 20 % (274 cases) going into the model test set (MTS). The split is stratified, and the process is repeated five times, where we split the whole dataset into training and testing sets randomly. The average results from the five repeated splits are reported. Please note that data division has been done from the first beginning of the machine learning pipeline, which prevents the information leakage issue in model testing.
- *Second*, the MDS data is preprocessed, as discussed in Section 2.3.1. The resulting data preprocessing operators are applied separately on the test set without refitting again on the test set.
- *Third*, a feature extraction step is conducted for the time-series data to calculate the statistically relevant features that best describe each time series. The resulting statistical features can be combined with the static features.

[6] *Alzheimer's Disease Neuroimaging Initiative (ADNI):* *https://www.adni.loni.usc.edu*.
[7] FreeSurfer neuroimaging toolkit: *https://surfer.nmr.mgh.harvard.edu/*.
[8] ADNI MRI acquisition procedure: *https://adni.loni.usc.edu/methods/mri-tool/mri-analysis/*.
[9] ADNI reference repository: *https://adni.bitbucket.io/reference/index.html*.

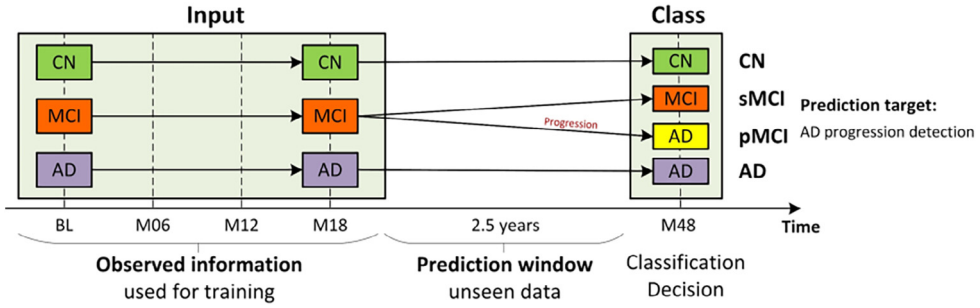

Fig. 1. Data formulation process.

Table 1
Descriptive statistics for ADNI dataset at baseline.

	CN (n = 419)	sMCI (n = 473)	pMCI (n = 140)	AD (n = 339)	Combined (n = 1371)
Gender (M/F)	191/228	283/190	86/54	187/152	747/624
Age (years)	73.84 ± 05.78	72.92 ± 07.76	73.89 ± 06.84	75.01 ± 07.81	73.82 ± 07.18
Education (years)	16.43 ± 02.70	15.80 ± 02.97	16.13 ± 02.71	15.13 ± 02.98	15.85 ± 02.90
FAQ	00.19 ± 00.73	02.10 ± 03.13	04.55 ± 04.54	13.32 ± 06.85	04.54 ± 06.65
MMSE	28.98 ± 01.14	27.63 ± 02.13	26.45 ± 02.09	21.94 ± 03.64	26.59 ± 03.63
MoCA	25.68 ± 01.97	23.14 ± 02.70	21.78 ± 01.99	17.48 ± 03.54	22.38 ± 04.08
APOE4	00.27 ± 00.48	00.51 ± 00.66	00.85 ± 00.71	00.85 ± 00.71	00.56 ± 00.67
ADAS-Cog 11	05.64 ± 02.83	09.13 ± 3.91	12.09 ± 03.66	19.64 ± 06.74	10.97 ± 06.99
ADAS-Cog 13	08.70 ± 04.09	14.80 ± 05.84	19.62 ± 05.05	30.00 ± 07.99	17.19 ± 10.03
RAVLT immediate	45.80 ± 09.72	36.41 ± 10.65	29.69 ± 07.08	22.64 ± 07.47	35.20 ± 12.80
RAVLT learn	06.03 ± 02.19	04.57 ± 02.50	03.06 ± 02.30	01.83 ± 01.77	04.19 ± 02.74
RAVLT forgetting	03.66 ± 02.73	04.42 ± 02.51	04.88 ± 02.19	04.45 ± 01.83	04.25 ± 02.44
RAVLT % forget	33.47 ± 26.89	53.70 ± 31.37	70.73 ± 28.79	89.44 ± 20.87	57.98 ± 34.71
CDR	00.084 ± 0.30	01.37 ± 00.86	02.07 ± 01.00	05.34 ± 02.21	01.96 ± 02.34

*Data are mean ± standard deviation.

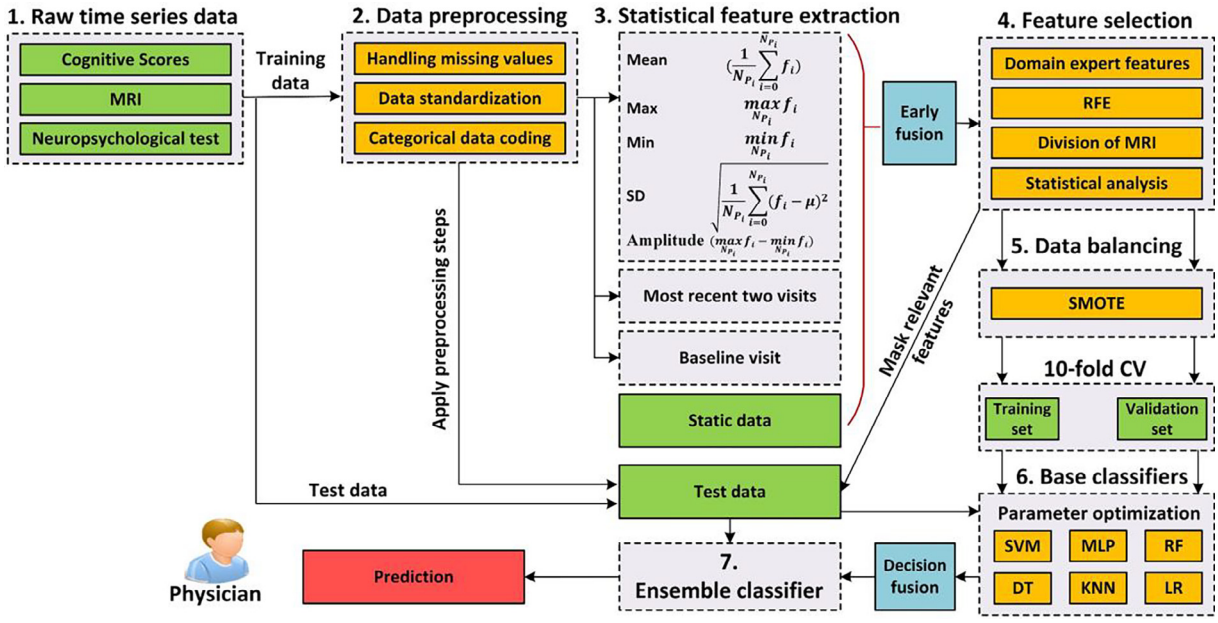

Fig. 2. Proposed AD progression prediction pipeline.

- *Fourth*, training the models with the most informative and discriminative list of features improves the model performance, enhances the model generalizability, and reduces the computation cost and complexity. A feature selection step is implemented prior to the training to identify the most relevant subset of features for our 4-class classification problem. The selected features are marked on the testing set to have the same features as those selected and used after the feature selection process.
- *Fifth*, the data over the four classes is unbalanced, and training with this data is likely to produce biased models [17]. The four classes CN, sMCI, pMCI, and AD, have proportions of 30.56 %,

34.50 %, 10.21 %, and 24.73 %, respectively. Namely, the percentage of pMCI is much smaller than CN, sMCI and AD. Therefore, we used the synthetic minority oversampling technique (SMOTE) to deal with the class imbalance in the MDS [49], but SMOTE is not applied to MTS to mimic real-world use. The MDS is further split into training and validation sets using the stratified 10-fold cross-validation (10-fold CV) technique. The resulting sets are used to train ML models to select the best list of features. The literature has confirmed that oversampling by SMOTE is one of the best methods for handling imbalanced datasets [50]. In addition, ML models achieve better performance using balanced datasets [50]. For these reasons, we decided not to train our ML models using imbalanced datasets.

- *Sixth*, MDS is used to train and validate the base ML classifiers based on the stratified 10-fold CV technique. The 10-fold CV ensures unbiased model fitting. The final evaluation of the models is performed using the reserved, unseen MTS. Six of the most popular and diverse ML models (i.e., SVM, RF, DT, XGB, KNN, and MLP) were selected to be trained individually on the MDS sets before all their possible combinations were tested on the MTS, this helped us choose which models were used to build a suitable ensemble classifier[check]. These techniques have been extensively used in the AD domain [17,20]. Models were trained on different modalities before their fusion. The lists of features that achieved the best performance from different modalities or combinations of modalities were used with their candidate classifiers to build the ensemble model.
- *Seventh*, we carried out the aggregation of base learners based on different fusion strategies in this step.

2.3.1. Data preprocessing

A) Missing data handling: Since the dataset features are numerical and categorical, the missing values were handled according to data type. For the baseline static data, we first removed any feature where more than 30 % were missing. Next, for those features with less than 30 % missing, we used the k-nearest neighbors (KNN) algorithm to impute missing values, where missing values were replaced using information from the cluster of subjects with the same diagnosis. That is, we found k neighbors, and then the imputed value was computed by averaging the values of those neighbors. In our study, k was set to 10 empirically; for the numerical values, the Euclidean distance was used; for categorical values, a distance of 0 was taken if both values were the same; otherwise, a distance of 1 was taken. For time-series data, we also removed any feature where more than 30 % were missing. Any patient cases with missing baseline readings were excluded from the study. Some critical features, such as CSF tau (83 %), were missing more than 30 % of the time series but were not missing baseline data. We opted to collect these features at baseline and consider them as background data (BG) by combining them with our static data. Moreover, the list of initial potential features was determined according to ADNI recommendations as well as AD diagnosis and progression literature [14]. Then a separate feature selection step was used to further collect significant features from this initial set. For handling non-existing time series values, we followed two sequential strategies according to the intuition of ADNI. First, we filled non-applicable values for every category of data according to ADNI procedures. For example, an ADNI 1 CN patient would not have an MRI scan at visit M18. As a result, we should not consider these types of values to be missing or consider that they are missing not at random. Many lab tests, cognitive tests, and neuroimaging scans are not done for specific diagnoses at specific visits. We followed an accurate procedure to fill these non-applicable values. If the diagnosis had not changed, we used forward filling with the previous values. If the diagnosis had changed, we considered the value as missing. This technique is common in the AD literature [47]. The second step is to determine the missing value from existing data using statistical or ML techniques. We opted for applying a medically intuitive and well-known method. Thus, in the case of numerical data, we used the mean value according to the different classes: CN, sMCI, pMCI, or AD. For categorical features, we used the mode value according to the patient class.

B) Data standardization: To ensure an equal level of importance, feature data were standardized using the z-score method, i.e., $z_j = (x_j - \mu_j)/\sigma_j$ where x_j is the participant's original value for feature j, z_j is the normalized value, μ_j is the feature's mean, and σ_j is the feature's standard deviation. The z-score method converts sets of data so they have a zero mean and unit standard deviation. The values of categorical features have been encoded.

2.3.2. Time series feature extraction

The intuition behind our approach is to transform time series data into a feature vector. The utilized time-series data have an extensive collection of features. If the time series data are simply flattened, the resulting feature vector will be quadrupled because we have four-time steps. Therefore, the resulting feature set will not be suitable due to the number of samples and overfitting is expected. Note that the length of each time step is six months. Our approach was to summarize the time series data without increasing the number of features. There are three main strategies to map time series data into feature vectors suitable for regular ML models.

Among strategies in the literature, this paper implements the following because it is the most popular method for converting time series data to feature vectors [51]. We aggregated features from the patient's four historical visits. We extracted time-domain statistical features only, this means we extract the statistical measurements that describe time series characteristics. For patient P_i, the aggregated feature $\bar{f}_{t_1,t_2,\cdots,t_T} = (\bar{f}_1, \bar{f}_2, \cdots, \bar{f}_A)$, where $\bar{f}_{t_1,t_2,\cdots,t_T}$ is a vector with A dimensions, and $\bar{f}_a$ for $a = 1, 2, \cdots, A$ could be the mean $(\frac{1}{N_{P_i}}\sum_{i=0}^{N_{P_i}} f_i)$, max $(\underset{N_{P_i}}{max} f_i)$, min $(\underset{N_{P_i}}{min} f_i)$, standard deviation $\sqrt{\frac{1}{N_{P_i}}\sum_{i=0}^{N_{P_i}}(f_i - \mu)^2}$, or amplitude $(\underset{N_{P_i}}{max} f_i - \underset{N_{P_i}}{min} f_i)$ features. Note with this strategy, the static data are later fused with the extracted features from time-series modalities.

2.3.3. Feature selection

Feature selection is a crucial step to find the most discriminative, informative, and concise subset of features from the available feature set. The selected features are expected to minimize redundancy, avoid overfitting, improve model generality, and reduce computational costs. Relevant features can be selected using filtering, wrapper, or domain expert approaches [52]. Each method has its own advantages and disadvantages [53]. Recommended features by domain experts could enhance model interpretability. Wrapper methods for training and testing classification models are commonly used to assess the importance of different subsets of features in the space of possible feature subsets [54]. These methods typically apply a greedy search algorithm to select the optimum subset of features based on the performance of a specific classification model such as SVM [55], Bagging [56], RF [57], or Naïve Bayes [56]. Filtering methods measure the importance of each feature separately using metrics like information gain or correlation coefficient.

We applied a collection of these methods to select the optimum sets of features from different modalities. More precisely, we utilized the information gain methodology (from the available filtering methods) to select the best features from the Static dataset. From the wrapper methods, we used the recursive feature

elimination (RFE) technique combined with RF (RFE-RF) and XGB (RFE-XGB) classifiers to measure feature importance. In the medical domain, studies asserted that RFE-RF is more efficient than other methods such as RFE-SVM to find highly discriminative and small feature sets [17]. We applied RFE-RF and RFE-XGB with 10-CV on the MDS. This 10-fold CV process was repeated 30 times using different stratified and random partitioning of the MDS to avoid any possible bias. RFE searches for the optimum combination of features that maximizes classifier performance using backward feature elimination based on a feature importance metric as the ranking measure. The selected features from all these methods were used to build the ensemble model. In addition, statistical analysis methods were used to analyze the selected features.

2.3.4. Ensemble classifier modeling

An ensemble of classifiers is a collection of representative classifiers that are accurate and diverse [35]. All classifiers are trained in tandem, and their decisions are combined according to some strategy. The enhanced performance of the ensemble is due to the ability to combine accurate predictions and correct errors from many diverse classifiers. Thus, getting a good accuracy/diversity trade-off becomes a challenging problem. Homogeneous ensembles (e.g., bagging, boosting, or random forest) utilize a single base classifier and achieve diversity by sampling from training data, assigning different weights to single data instances and/or sampling from the feature set. Heterogeneous ensembles achieve diversity using a pool of distinct base classifiers (e.g., SVM, RF, DT, MLP, etc.). They usually include a form of *meta*-learning such as techniques called stacking, voting, or ensemble selection. The ensemble selection method includes the following three main steps:

- *Base model generation*: In this step, different models were trained with the data. More precisely, we considered RF, SVM, KNN, XGB, MLP, and DT, as the ML techniques in this study. It is worth noting that even if other ML algorithms could have been used, we chose these ML techniques due to their wide use in the medical domain [58] and to their potentially diverse and complementary predictions.
- *Ensemble pruning*: In this step, a subset of the models previously generated were selected to become members of the ensemble using the accuracy/diversity trade-off as the primary criterion. Maximizing accuracy and diversity produce enhanced stability as a side effect [59]. Diversity should not depend on the classifiers' outputs but on how they differ and become complementary [46]. General measures like variance and mutual information are not accurate here because they simply measure how members are diverse without employing any information on the correctness of the output [60]. We utilized the so-called "*binary oracle method*", which knows the correct answers and evaluates diversity based on whether each classifier is correct or incorrect [61]. Formally, if we take an ensemble $H=\{c_t, t=1,2\cdots C\}$ of C classifiers to classify $D=\{(x_1,y_1),(x_2,y_2)\cdots(x_N,y_N)\}$ of N samples, $y_i\in\Omega=\{l_1,l_2\cdots l_m\}$ for m class labels, $x_j\in\mathbb{R}^n$; a classifier c_i assigns a class label from Ω to an input example x_j, i.e. $c_i:\mathbb{R}^n\rightarrow\Omega$, or $c_i(x_j)\in\Omega$, $i=1,2\cdots C$ is the output of i th classifier c_i for the j th sample x_j, and $r(x_j)$ is the true label for $j=1,2,\cdots N$. For classifiers c_1 and c_2, the correctness of classification can be collected in two binary vectors $v^{c_1}\in\mathbb{R}^N$ and $v^{c_2}\in\mathbb{R}^N$. For a two-classifier evaluation, N^{11} denotes the number of times c_1 and c_2 are correct, N^{00} the number of times c_1 and c_2 are incorrect, and N^{10} and N^{01} the number of times where just one of c_1 and c_2 is correct, respectively, and where N=N^{11} +N^{10}+N^{01}+N^{00}. For more than two classifiers, diversity is calculated as the mean of the pairwise values. Single measures of diversity are inadequate [46], so we utilized three measures to test diversity.

The first measure is the *correlation* between the errors calculated by Equation 1, where $Cov(.)$ and $Var(.)$ are the covariance and variance, respectively. The goal is to minimize the mean pairwise ρ s.

$$\rho_{c_1,c_2}=\frac{Cov(v^{c_1},v^{c_2})}{\sqrt{Var(v^{c_1})Var(v^{c_2})}} \tag{1}$$

The second measure is the *Q statistic* [62], as defined in Equation 2. The goal is to minimize the mean pairwise Q s.

$$Q_{c_1,c_2}=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}} \tag{2}$$

$Q_{c_1,c_2}\in[-1,1]$, and 0 means independent classifiers.

The third measure is the *disagreement* measure [63] calculated by Equation 3. The goal is to maximize the mean pairwise D s.

$$D_{c_1,c_2}=\frac{N^{01}+N^{10}}{\mathrm{N}} \tag{3}$$

- *Ensemble integration*: In this step, we pay attention to how to combine single decisions into a unique agreed ensemble decision. The strategy to apply depends on the type of classifier outputs. For each instance x_j, each classifier c_i outputs discriminant measure $o_i=c_i(x_j)$, where $o_i\in\{l_1,l_2\cdots l_m\}$ or $o_i=[c_{i,1}(x_j),c_{i,2}(x_j)\cdots c_{i,m}(x_j)]^T$ and $c_{i,k}(x_j)$ is the predicted probability (or the degree of support) by a classifier c_i for class l_k using sample x_j. By combining all classifiers, the decisions are $O=[o_1,o_2\cdots o_N]^T$, or more generally, as shown in Equation 4.

$$O=\begin{bmatrix} c_{1,1}(x_j) & \cdots & c_{1,k}(x_j) & \cdots & c_{1,m}(x_j)\\ \cdots & & & & \\ c_{i,1}(x_j) & \cdots & c_{i,k}(x_j) & \cdots & c_{i,m}(x_j)\\ \cdots & & & & \\ c_{C,1}(x_j) & \cdots & c_{C,k}(x_j) & \cdots & c_{C,m}(x_j)\end{bmatrix} \tag{4}$$

The final decision of the classifier ensemble is a combination of the base learners' decisions $H(x_j)=F(c_1(x_j),c_2(x_j),\cdots c_N(x_j))=F(O)=F(o_1,o_2\cdots o_N)$. When classifiers return class labels, F includes: (i) the linear fusion of members' decisions via majority voting, i.e., $H(x_j)=\underset{k=1,2\cdots m}{\operatorname{argmax}}\sum_{t=1}^{C}c_{t,k}(x_j)$; (ii) weighted majority voting $H(x_j)=\underset{k=1,2\cdots m}{\operatorname{argmax}}\sum_{t=1}^{C}w_t c_{t,k}(x_j)$; (iii) learning based non-linear fusion (e.g., stacking); or (iv) mixture of experts [64]. When classifiers return the degree of certainty, the posterior probability can be calculated as $P(c_i|x_j)=o_i$ and $c_{i,k}(x_j)\in o_i$ is defined as before. Without loss of generality, we can let o_i be normalized, i.e., $o_i\in$; the combination of classifiers is the posterior probability of $P(l_k|c_1(x_j),c_2(x_j),\cdots c_N(x_j))$. Decision rules are defined as follows: (i) product rule, $\max_{k=1\cdots m}\prod_{i=1}^{C}c_{i,k}(x_j)$; (ii) sum rule, $\max_{k=1\cdots m}\sum_{i=1}^{C}c_{i,k}(x_j)$; (iii) max rule, $\max_{k=1\cdots m}\max_{i=1\cdots C}c_{i,k}(x_j)$; (iv) min rule, $\max_{k=1\cdots m}\min_{i=1\cdots C}c_{i,k}(x_j)$; or (v) median rule, $\max_{k=1\cdots m}\frac{1}{C}\sum_{i=1}^{C}c_{i,k}(x_j)$. In stacking, a meta classifier is used to learn the aggregation function using the base classifiers' outputs and makes the final decision.

2.3.5. Implementation details and performance metrics

We used *Python* 3.7, *Scikit-Learn* 0.21.3, *mlxtend* 0.17.0, and *XGBoost* 0.90 to generate base classifiers and to build the ensemble models. The performance of the base classifiers and ensemble models was measured using the standard five metrics of accuracy, precision, recall, F1-score, and balanced accuracy (BA), which are defined as follows in Equations 5–9, respectively:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad 7)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (9)$$

3. Results and discussion

Designing an accurate and stable ensemble classifier is a challenging task. In our list of experiments, we gradually test different dimensions and tuning methods, and select the optimum path to reach the best design, which is more accurate and medically relevant, as shown in Fig. 3.

1. First, we select the most critical features using the information gain and recursive feature elimination techniques from every individual modality. This step tried to select the best list of features to make the model more accurate and less expensive medically.
2. Having the list of features from each modality, we need to select the best design for the ensemble model. Note that it is well-known in the literature that ensemble classifiers achieve better results than all its base classifiers. We followed two paths to optimize the ensemble design process:
   a. The first path is to fuse all selected features in step 1 (i.e., early fusion strategy). The main idea here is to explore the possible role of the hidden relationships among features of different modalities to improve the model performance. We used the collected list of features to compare the homogeneous classifiers-based methods like bagging and boosting with the heterogeneous ensemble techniques like voting and stacking. The bagging and boosting methods are based on the most accurate base classifier, so we optimize a set of base classifiers, including KNN, decision tree, SVM, etc., to select the best performing technique. The voting and stacking techniques are based on accuracy/diversity trade-off explorations, so we measure both diversity (using Q-statistic, correlation coefficient, and disagreement) and performance using accuracy, precision, recall, and f1 score. To select the best combination of diverse and accurate models, we test many options. The main reason for combining base classifiers is to choose the best number of base classifiers that are most accurate and most diverse, which achieve the best ensemble performance.
   b. The second path is to build a heterogeneous classifier for each different modality and fuse the decisions of these base classifiers using ensemble techniques like stacking or voting (i.e., late fusion). For each modality, we trained six classifiers and selected the best one or a set of most accurate classifiers. In addition, the possible roles are explored for the different combinations of modalities. The role of 13 combinations of AD modalities is evaluated, and each dataset is tested using six machine learning algorithms. This results in 78 base classifiers. Here, we explored the role of the medical categorization of features and used each base classifier as an expert in

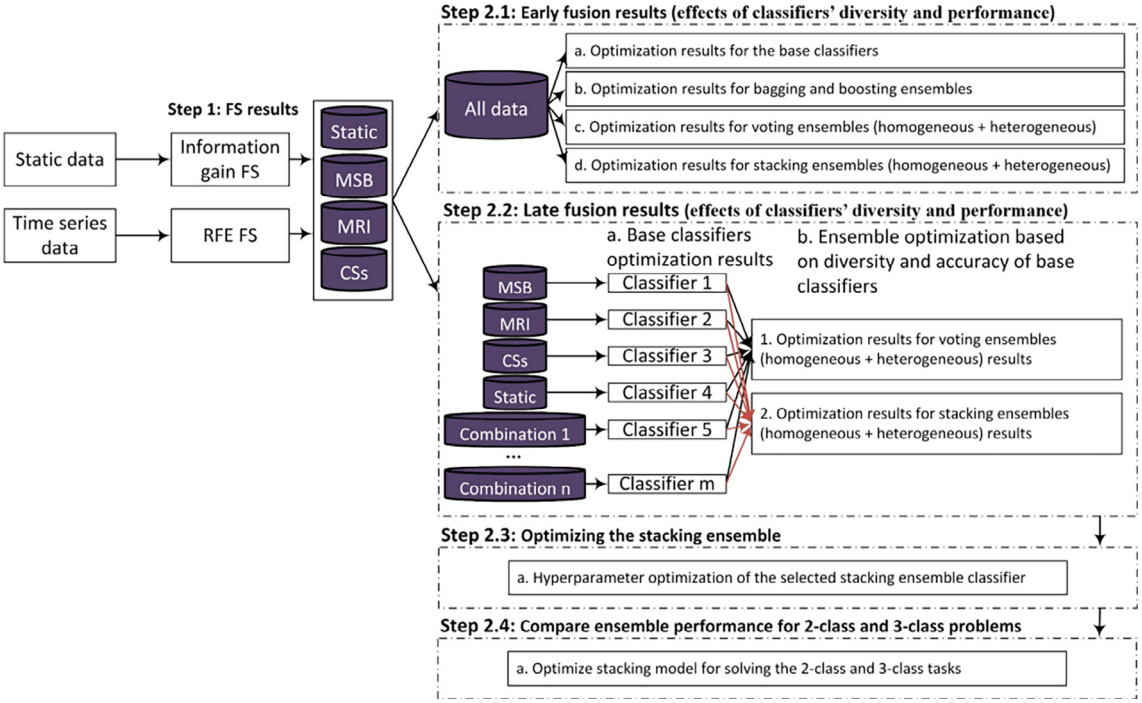

Fig. 3. Results roadmap.

a specific domain. For example, the MRI-based classifier is like an expert in making progression detection decisions based on neuroimages.

c. After selecting the best performing ensemble model based on the accuracy and diversity of the base classifiers, we optimized the hyperparameters of the selected stacking ensemble to further enhance its performance.

d. Finally, we tested the performance of the optimized ensemble to solve the 2-class and 3-class classification tasks for Alzheimer's disease prediction.

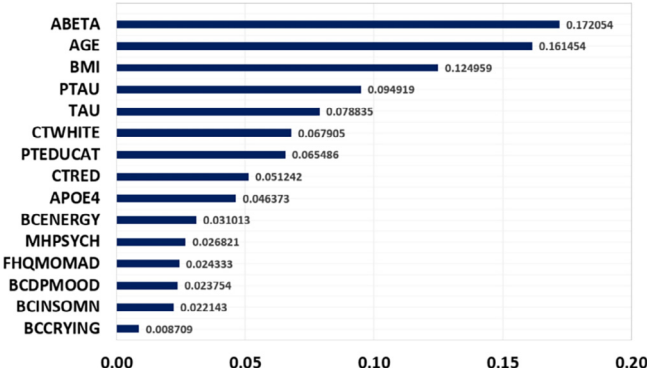

Fig. 4. The importance of static features.

3.1. Feature analysis

To train the base models, we have two choices. First, all data modalities can be fused, and a random subset of the features is selected to train each base classifier. Second, it is possible to benefit from the medical relationships among features of each modality and independently train base models using every single modality. The second choice is called the decision fusion paradigm. In this section, we explain how to select the most discriminative features from each modality. These features are used either in the early or late fusion designs.

3.1.1. Static feature analysis

To select the best set of Static features, we applied the information gain filter-based approach. We selected the top 15 (~24 %) discriminative features, including ABETA, AGE, APOE4, BCCRYING, BCDPMOOD, BCENERGY, BCINSOMN, BMI, CTRED, CTWHITE, TAU, FHQMOMAD, MHPSYCH, PTAU, and PTEDUCAT (see the Supplementary File 2 for further details). The values of these features are statistically significantly different between the four diagnosis groups ($P = 0.05$). Fig. 4 shows the feature importance based on the information gain.

3.1.2. Time series feature analysis

To select the best features from the time-series modalities, we followed a three steps methodology. First, we selected the candi-

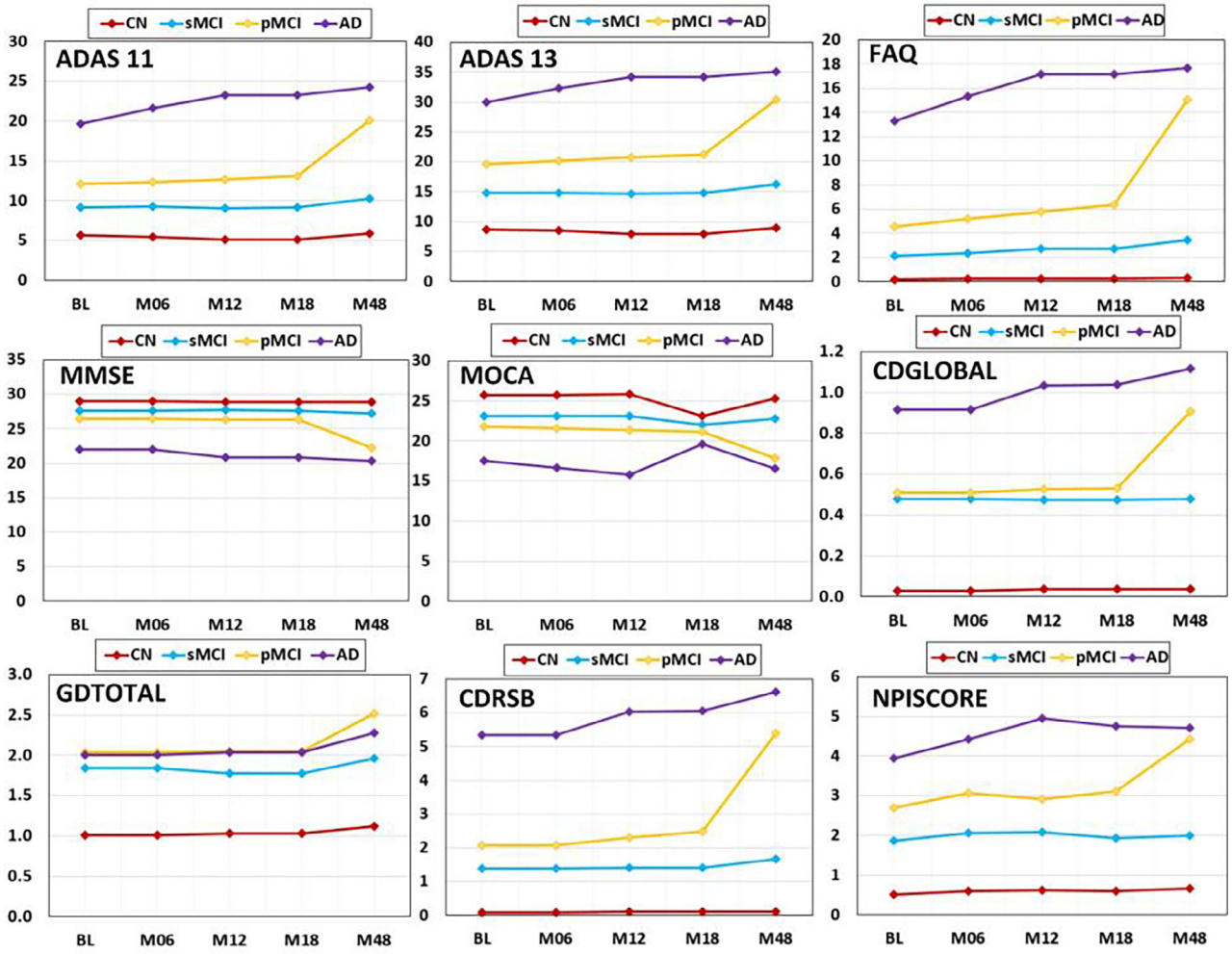

Fig. 5. Cognitive scores temporal changes.

date covariates from each modality according to their perceived clinical importance and to our statistical analysis. Second, we extracted statistical features for each modality, as proposed in Section 2. Third, we used feature selection techniques to automatically select the most informative and discriminative features for AD progression detection. It is worth noting that a detailed description of all the features in this paper is provided in Supplementary File 2.

3.1.2.1. Relevant features collection. Regarding CS features: based on the feature selection step, nine features (CSs) are used in this study. As shown in Fig. 5, these scores let us discriminate the four classes very well at each visit. For example, the AD class has the highest values for ADAS 11, ADAS 13, CDRSB, NPISCORE, CDGLOBAL, and FAQ, followed by pMCI, sMCI, and then CN. On the other hand, the AD class has the lowest MMSE and MOCA, followed by pMCI, sMCI, and then CN. For GDTOTAL, both AD and pMCI have comparable values, but sMCI has higher values than CN. In addition, it is of note that the pMCI class has statistically significantly different values collected in visits at M18 compared to at M48 ($P < 0.05$). Collectively, the box plots in Fig. 6(1) illustrate that the four classes have significantly different values for cognitive scores in the BL, M06, M12, and M18 data ($P < 0.001$). In addition, Fig. 6(2) asserts the role of cognitive scores to clearly separate the four classes ($P < 0.05$).

Regarding NSB features: we selected 17 features including RAVLT and ADNI MEM among others (see Supplementary File 2). The selected features are medically well known and extensively used in the literature [1]. Due to space restrictions, Fig. 7 shows box plots for only nine of these features based on the BL, M06, M12, and M18 data. As can be seen, all features are statistically significantly different between the four classes ($P < 0.01$). Furthermore, these features discriminate the classes in the same manner at M48.

Regarding MRI features: recent studies have asserted that AD progression is tightly correlated with atrophies in the structures of the medial temporal lobe (MTL) [25,65,66]. The MTL includes a set of anatomical regions such as the hippocampus, amygdala, entorhinal, and parahippocampal cortices. Each of these regions has left and right parts. In addition, cortical thickness has a high predictive value for AD progression detection [67,68]. We collected the volume, surface area, and cortical thickness for each of these parts. Out of the 312 MRI features, we selected 43 features related to volume and cortical thickness of the left part and suitable structures, including HIPPOCAMPUS, AMYGDALA, PARAHIPPOCAMPAL, ENTORHINAL, VENTRICLES, FUSIFORM, INFERIOR/MIDDLE/ SUPERIOR TEMPORAL, and INSULA. These features have the highest information gain values. The weights of these features according to the information gain can be found in Supplementary File 2, Fig. S1. The most important features are the left and right volumes of the hippocampus and the cortical thickness of the left entorhinal. Fig. 8 illustrates the top 10 volumes (left part) and top 10 cortical thicknesses (right part). As can be seen, the selected features can clearly discriminate different classes, especially CN vs AD ($P < 0.05$). The discriminative power of the selected biomarkers is on par with results from the literature [1,69,70–72].

3.1.2.2. Statistical features extraction. In this step, we convert time series data into a suitable format for conventional ML models. As discussed in Section 2, we have two strategies to extract features from time-series data, including baseline data and statistical data. Statistical features are collected from the four readings (BL, M06, M12, and M18) associated with each patient. The calculated fea-

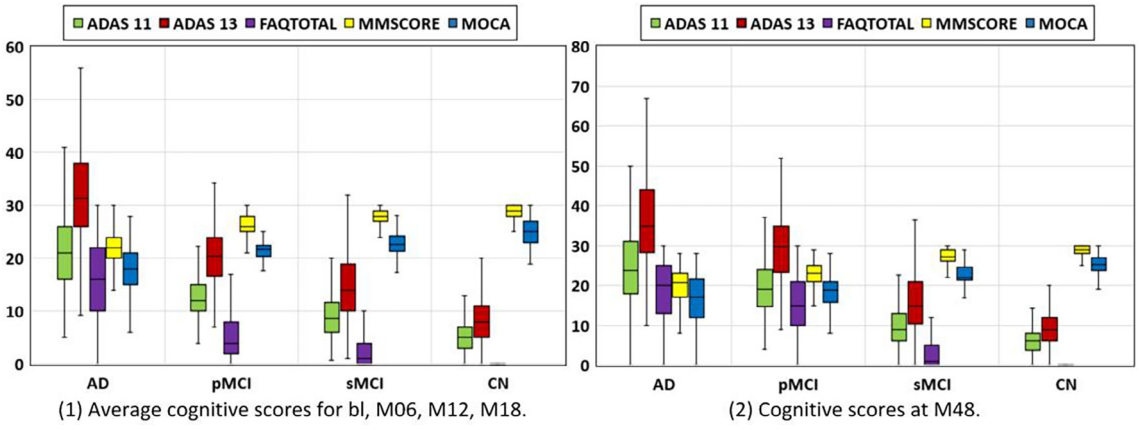

Fig. 6. Cognitive scores discriminative features.

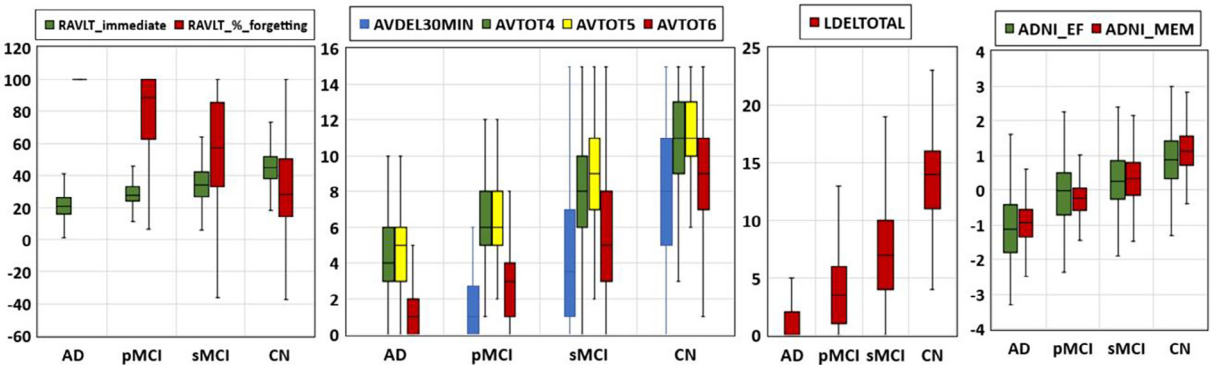

Fig. 7. Neuropsychological battery discriminative features.

tures include the maximum, minimum, average, and standard deviation. These feature sets are made up of 36, 68, and 172 features from CSs, NSB, and MRI, respectively.

3.1.2.3. Time series feature selection. As discussed in the previous section, we extracted many features from every modality; however, not all of these are discriminative and helpful in AD progression detection. Selecting relevant features improves model performance and simplicity, reduces measurement costs, and enhances model understandability [73]. As a result, a feature selection step is crucial. Feature selection was based on the pre-defined model development set (MDS) (80 % of the entire dataset). The ten fitted and tested RFE-RF models selected 20 features (55.5 % of the total features) for the CS modality, and the RFE-XGB models selected 25 features (69.4 % of the entire feature set). We found 18 features recommended by both types of models. The final set of features consists of the union of the sets provided by both methods (i.e., 27 features). Regarding the MRI modality, RFE-RF recommended 64 features, and RFE-XGB recommended 29 features. We found that RFE-RF also recommended 65.5 % of the RFE-XGB features. We used the combination of these features (i.e., 74 features, 43 % of the total feature set). For the NSB modality, RFE-RF selected 16 features, and RFE-XGB selected 12 features. We found 11 features recommended by both methods. We combined both sets and obtained 25 features (36.8 % of the complete NSB set). The selected features are ranged from maximum, minimum, standard deviation, and average. A full description of these features can be found in Supplementary File 2. We reevaluated the feature weights using the information gain filter method to confirm that the selected features are statistically significant. We found the selected features had the highest weights for all modalities.

3.2. Design of ensemble classifiers

After selecting the best features associated in each modality, we implemented a set of experiments to choose the best design scheme for the AD detection ensemble model. The first set of experiments in Section 3.2.1 fuses all modalities into one source (early fusion) and uses the resulting set to build an ensemble model. In this design, we depend on the hidden relationships among features of different modalities. The second set of experiments, discussed in Section 3.2.2, builds separate classifiers for every modality and their different integrations and fuses the decision of all base classifiers to make the final decision (i.e., late fusion). In the second design method, we independently utilized the medical relationships among features in each data source by building a separate base classifier for each data type.

3.2.1. Early fusion-based ensemble modeling

The fused data set has 141 features from the four modalities. The first experiment compared the homogeneous methods of bagging and boosting with the heterogeneous ensemble techniques of voting and stacking. On the one hand, the bagging and boosting methods are based on the most accurate base classifier. On the other hand, the accuracy/diversity trade-off is examined for the voting and stacking methods (see Fig. 9). In other words, we built the models with the most accurate base classifiers and with the most diverse base classifiers to assess the effect of both measures (i.e., accuracy and diversity) on the ensemble performance. The individual rate is the minimum error rate of a classifier, as shown in equation 5, this is used as a performance metric in addition to the other metrics of precision, recall, and F1-score.

$$Minimum\ \ error\ \ rate = \min_{k}\left(\frac{1}{m}\sum_{j=1}^{m} e_k(j)\right) \tag{5}$$

where $e_k(j)$ is the classification error of classifier k for data j. We examined various diversity measures. This experiment was aimed at finding the optimum set of classifiers with the highest performance and diversity.

Base classifiers were fitted using the MDS with the stratified 10-CV technique. To get more stable results, each experiment was repeated 30 times and the averaged results are reported here. Then, the performance of the model was measured using the unseen MTS test set, as shown in Table 2. For a fairer comparison,

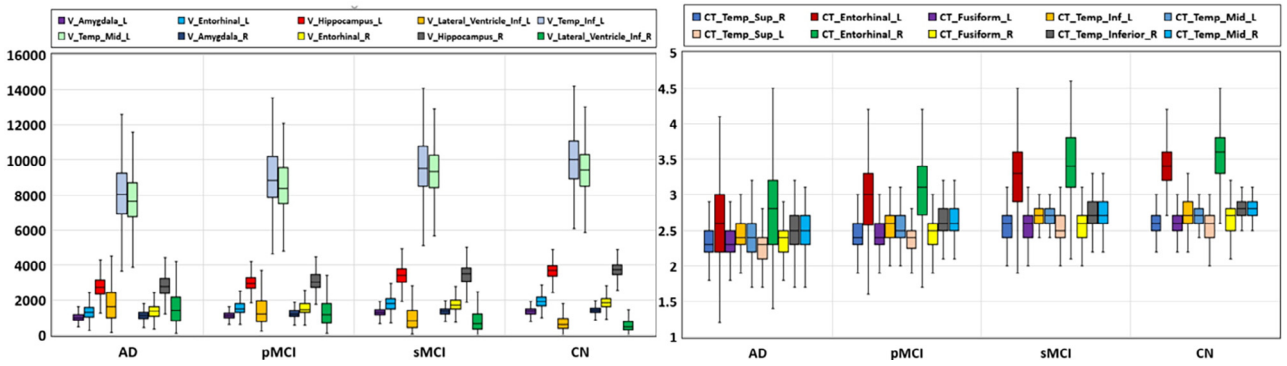

Fig. 8. Volume and cortical thickness of the most important MRI features.

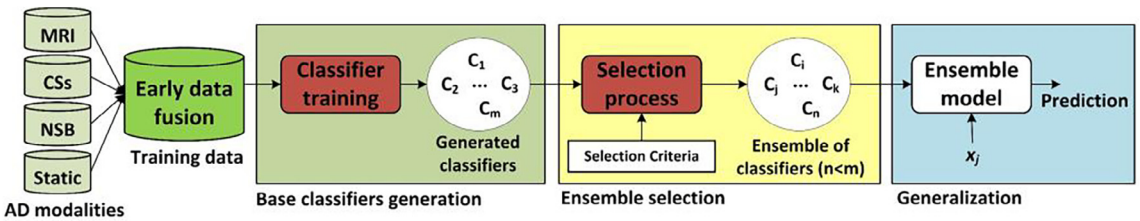

Fig. 9. Early fusion-based ensemble selection.

Table 2
Validation results for the base classifiers.

Classifier index	Base classifier	Performance measurement (%)					Error rate	Classifier rank	Diversity measurement		
		Accuracy	Precision	Recall	F1-score	BA			Q-statistic	Correlation	Disagreement
1	**RF**	**84.95**	**84.88**	**84.95**	**84.87**	**79.38**	**15.05**	**1**	0.7310	0.4322	0.1181
2	SVM	83.98	84.98	83.98	84.38	80.59	16.02	3	**0.7554**	**0.4816**	**0.1052**
3	KNN	74.51	83.77	74.51	74.48	78.43	25.49	6	**0.6253**	**0.3090**	**0.1812**
4	XGB	84.46	84.07	84.47	84.25	77.37	15.54	2	0.7438	0.4671	0.1092
5	DT	77.43	78.21	77.43	77.66	72.93	23.06	5	0.6797	0.3627	0.1505
6	MLP	82.28	82.89	82.28	82.57	75.89	17.72	4	0.7338	0.4645	0.1141

the base classifiers were trained using the default parameters of Scikit-learn [74]. It is of note that the RF classifier achieved the highest results (Accuracy = 84.95 %, Precision = 84.88 %, Recall = 84.95 %, F1-score = 84.87 %, and BA = 79.38 %), while KNN had the worst results (Accuracy = 74.51 %, Precision = 83.77 %, Recall = 74.51 %, F1-score = 74.48 %, and BA = 78.43 %). Both XGB and RF are ensemble models, and, as expected, they achieved the best two results.

Table 2 also shows the average of the pairwise diversities for each base classifier. We calculated diversity using three popular techniques: Q-statistic, correlation coefficient, and disagreement. It should be noted that all these diversity measures sorted the base classifiers in the same order. KNN is the most diverse classifier, but it has the worst performance while SVM has the lowest diversity. To study the effect of the diversity/accuracy trade-off on classifier subset selection, we performed experiments using voting and stacking ensemble techniques. We used soft voting because it achieves better results than hard voting. As selecting the optimum subset of classifiers requires an exhaustive search among $2^C + 1$ possible combinations, we used a simple greedy method to select the optimum (i.e., the combination with the best balance between diversity and accuracy) subset of base classifiers that work well together, we started with an ensemble size of $n = 2$. The results were compared with the bagging and AdaBoost techniques using RF (i.e., the most accurate classifier) as the base estimator. Bagging and AdaBoost were expected to generate diverse classifiers [46].

Regarding the voting technique, Table 3 shows how just putting together all the base classifiers does not achieve the best performance. These results are most likely because of the insufficient diversity and poor performance in the learner pool. RF and XGB are the two most accurate classifiers, and together they achieve an accuracy of 85.44 %, precision of 85.00 %, recall of 85.44 %, F1-score of 85.19 %, and BA of 78.92 %. However, using the two most diverse techniques (i.e., KNN and DT), we achieved an accuracy of 78.16 %, precision of 81.55 %, recall of 78.16 %, F1-score of 78.95 %, and BA of 74.96 %. As confirmed by the literature, paying attention to only diversity is likely to produce poorer performance [75 76].

Accordingly, diversity must be balanced with accuracy to create a well-performing ensemble [46]. Using the most accurate three classifiers (i.e., RF, XGB, and SVM), we achieved an accuracy of 85.19 %, precision of 80.99 %, recall of 80.99 %, F1-score of 80.99 %, and BA of 78.74 %. However, this ensemble does not improve on the performance of RF and XGB. This is because SVM has the lowest diversity level and so did not add value to the ensemble. This conclusion is confirmed when using RF and XGB with KNN. Although KNN is less accurate than SVM, it adds more accuracy to the ensemble. This is because KNN has the highest diversity level. The combination of RF, XGB, and KNN achieved the best results (i.e., Accuracy = 86.17 %, Precision = 88.10 %, Recall = 86.17 %, F1-score = 86.67 %, and BA = 84.41 %). By repeating the most accurate classifier (i.e., RF) twice, the ensemble achieved better results (i.e., Accuracy = 86.65 %, Precision = 88.16 %, Recall = 86.65 %, F1-score = 87.06, and BA = 84.81). By repeating the RF clas-

Table 3
The effect of diversity and performance on voting and stacking ensembles.

Base classifiers selection strategy	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	BA (%)
Vote: all models	RF, SVM, KNN, XGB, DT, MLP	84.71	85.22	84.71	84.90	79.16
Vote: 2 most accurate	RF, XGB	85.44	85.00	85.44	85.19	78.92
Vote: 2 most diverse	KNN, DT	78.16	81.55	78.16	78.95	74.96
Vote: 1 most accurate + 1 most diverse	RF + KNN	78.88	84.60	78.88	79.75	80.06
Vote: 3 most accurate	RF, XGB, SVM	85.19	80.99	80.99	80.99	78.74
Vote: 3 most diverse	KNN, DT, MLP	81.80	84.20	81.80	82.46	78.80
Vote: 2 most accurate + 1 most diverse	**RF, XGB, KNN**	**86.17**	**88.10**	**86.17**	**86.67**	**84.41**
Vote: 2 most accurate + 2 most diverse	RF, XGB, KNN, DT	83.74	85.00	83.74	84.14	79.29
Vote: (2 + 1) most accurate + 1 most diverse	RF, XGB, KNN, RF	86.65	88.16	86.65	87.06	84.81
Vote: 2 most accurate + 2nd most diverse	RF, XGB, DT	85.19	84.78	85.19	84.91	79.23
Stack: all models	RF, XGB, SVM, MLP, KNN, DT	85.29	84.59	85.29	84.91	77.92
Stack: remove worst 1	RF, XGB, SVM, DT, MLP	84.04	83.98	84.04	84.00	77.35
Stack: remove worst 2	RF, XGB, SVM, MLP	85.02	84.20	85.02	84.56	77.28
Stack: remove worst 3	RF, XGB, SVM	84.16	84.02	84.16	84.08	77.44
Stack: 2 most accurate	RF, XGB	84.93	84.17	84.93	84.51	77.32
Stack: remove 1 lowest diversity	RF, XGB, MLP, DT, KNN	83.60	83.37	83.60	83.46	77.00
Stack: remove 2 lowest diversity	RF, MLP, DT, KNN	83.33	83.65	83.33	83.44	77.71
Stack: remove 3 lowest diversity	RF, DT, KNN	80.30	80.91	80.30	80.53	74.50
Stack: remove 4 lowest diversity	DT, KNN	79.89	80.65	79.89	80.19	74.23
Stack: 2 most accurate + 1 most diverse	**RF, XGB, KNN**	**85.51**	**85.13**	**85.51**	**85.29**	**78.84**
Stack: 2 most accurate + 2 most diverse	RF, XGB, KNN, DT	83.40	83.33	83.40	83.34	76.92
AdaBoost: 1 most accurate	RF	83.50	82.19	83.50	82.02	72.56
AdaBoost: 1 diverse	DT	79.61	80.42	79.61	79.92	74.02
Bagging: 1 most accurate	**RF**	**85.75**	**86.24**	**85.75**	**85.95**	**81.77**
Bagging: 1 most diverse	KNN	74.48	83.36	74.48	74.49	78.19

sifier experiment, we added more weight to this classifier. However, repeating RF more than twice had no effect on the performance. On the other hand, repeating the KNN classifier did not enhance the classifier performance. We checked the different combinations, and none achieved better results. Our vote ensembles give the same weight to all base classifiers. Nevertheless, some base classifiers can be more important than others. Alternatively, the stacking ensemble uses a meta classifier to learn the optimum combination of base classifiers.

Regarding the stacking technique: we selected logistic regression (LR) to play the role of the meta classifier. The utilized decisions of the base classifiers are represented in the form of probability distributions. For example, for a specific training example x_j, in a 3-class setting $[l_1, l_2, l_3]$ with two level-0 classifiers $\{c_1, c_2\}$, these base classifiers might make probability predictions of $c_1 : [0.2, 0.5, 0.3]$ and $c_2 : [0.3, 0.4, 0.4]$. The level-1 meta classifier receives $x_j = [0.2, 0.5, 0.3, 0.3, 0.4, 0.4]$. The most two accurate base classifiers achieve performance of: Accuracy = 84.93 %, Precision = 84.17 %, Recall = 84.93 %, F1-score = 84.51 %, and BA = 77.32 %. However, the two most diverse classifiers achieve performance of: Accuracy = 79.89 %, Precision = 80.65 %, Recall = 79.89 %, F1-score = 80.19 %, and BA = 74.23 %. The stacked ensemble shows the same behavior as the voting ensemble method. If we concentrate on the accuracy of the models, then removing the worst classifiers does not significantly enhance the performance. The same for diversity, where removing the least diverse classifiers sometimes degrades the performance if the removed classifiers are highly accurate. Thus, the best solution is the one yielding the best balance between accuracy and diversity. This is proved when combining the most accurate classifiers (RF + XGB) and the most diverse classifier (KNN). In this case, the ensemble achieves the best performance: Accuracy = 85.51 %, Precision = 85.13 %, Recall = 85.51 %, F1-score = 85.29 %, and BA = 78.84 %. The main difference is that the stacking method achieved better performance than voting when we used all base classifiers (i.e., Accuracy = 85.29 %, Precision = 84.59 %, Recall = 85.29 %, F1-score = 84.91 %, and BA = 77.92 %). We note that adding such diversity to the ensemble does not improve performance. This means that diversity enforces regularization for the ensemble because the performance stops improving when there is no more diversity to extract from the pool of base classifiers [75,61].

Both AdaBoost and Bagging ensemble techniques were checked using the most accurate and diverse models. AdaBoost does not achieve good performance using RF and DT as the base classifiers. Bagging on the other hand achieves higher performance than stacking using RF as the base classifier (i.e., Accuracy = 85.75 %, Precision = 86.24 %, Recall = 85.75 %, F1-score = 85.95 %, and BA = 81.77 %). Note that all other parameters for both AdaBoost and Bagging were set to the default Scikit-Learn values. We conclude that the performance of the base classifiers has much more effect on the ensemble performance than diversity.

3.2.2. Late fusion-based ensemble modeling

Building an ensemble based on a set of heterogeneous classifiers trained using distinct feature sets often produces an accurate model because by considering heterogeneous base classifiers, they are likely to have complementary decisions and uncorrelated errors [36]. In this experiment, we exploited the medical relationships among the features of each of our four modalities by training a separate classifier for each one. In addition, we tested different combinations of modalities to check the effect of removing noise in different datasets on classifier performance. We implemented six classifiers for each modality and for their different combinations. The selection criteria for individual classifiers were mainly based on their performance and diversity.

For each dataset, we selected the optimum classifier or set of classifiers. Fig. 10 illustrates the detailed process for creating a customized ensemble model, where an exhaustive search is used to find the best representative models for each data type. It is worth noting that because the number of options is not significant when checking for the best model, we discarded here the use of heuristic optimization techniques such as multi-objective optimization techniques. First, we trained the six ML models with every modality and different combinations of modalities. We used the default hyperparameter settings from Scikit-Learn for all ML models. Second, we selected the optimum set of base classifiers, which are the most accurate and diverse. Third, the final ensemble decision is computed as the fusion of the single decisions made by the selected classifiers. We tested both voting and stacking as fusion strategies.

We evaluated the role of 13 combinations of AD modalities, and each dataset was tested using the six ML algorithms under study. This resulted in 78 base classifiers. Table 4 shows the testing results of these classifiers using the MTS data. It should be noted that the XGB classifier achieves the best performance in all experiments. In addition, the CSs turned out to have the highest performance compared with other single modalities. As a result, its combination with other modalities usually improves the performance. The optimum results were achieved by excluding the MRI data, i.e., the combination of CSs, NSB, and Static (Accuracy = 85.92 %, Precision = 86.95 %, Recall = 85.92 %, F1-score = 86.37 %, and BA = 81.11 %). In addition, the fusion of CSs and NSB achieves results not far from those reported by the above CSs, NSB, and Static combination (Accuracy = 85.68 %, Precision = 86.37 %, Recall = 85.68 %, F1-score = 86 %, and BA = 80.17 %).

Notice that, the addition of Static modalities improves the performance (Accuracy + 0.24 %, Precision + 0.58 %, Recall + 0.24 %, F1-score + 0.37 %, and BA + 0.94 %). The average of pairwise diversities for the base classifiers is shown in Table 4. Diversity was calculated among the classifiers of every dataset and the 78 classifiers. Both performance and diversity of the base classifiers were used to select ensemble members. Note that the same base classifier and/or the same dataset could be selected more than once. We

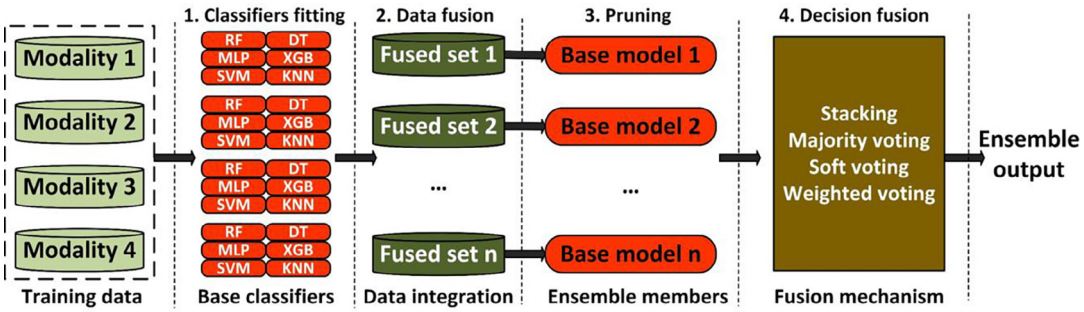

Fig. 10. Ensemble model construction steps.

Table 4
The performance of the base models for different modalities and their combinations.

Modality	#	Model	Performance (%)					Overall Diversity			Diversity of each data set		
			Acc	Pre	Rec	F1	BA	Q.	Corr.	Dis.	Q.	Corr.	Dis.
NSB	1	RF	75.96	76.22	75.96	75.89	71.26	0.3175	0.1390	0.3497	0.7272	0.4523	0.1335
	2	SVM	71.84	74.17	71.84	72.09	70.59	0.2866	0.1333	0.3533	0.7544	0.4983	0.1181
	3	**XGB (10)**	**77.18**	**78.11**	**77.18**	**77.51**	**72.53**	0.3018	0.1376	0.3537	0.7544	0.4969	0.1197
	4	KNN	68.93	72.60	68.93	68.73	68.01	0.2626	0.1249	0.3731	0.6769	0.3926	0.1691
	5	DT	74.10	74.67	74.10	74.15	70.29	0.2487	0.1087	0.3743	**0.6345**	**0.3426**	**0.1812**
	6	MLP	73.99	75.20	73.99	74.43	70.68	0.2846	0.1286	0.3621	0.7167	0.4362	0.1440
CSs	7	RF	78.77	80.24	78.77	79.24	73.92	0.3760	0.1691	0.3359	0.7959	0.5652	0.0858
	8	SVM	76.94	83.05	76.94	78.38	76.31	0.3227	0.1497	0.3465	0.7906	0.5616	0.0890
	9	**XGB (6)**	**80.83**	**81.95**	**80.83**	**81.29**	**75.93**	0.3329	0.1445	0.3411	0.7931	0.5455	0.0906
	10	KNN	74.51	81.78	74.51	76.29	73.90	0.3420	0.1612	0.3471	0.7832	0.5395	0.0995
	11	DT	77.64	79.31	77.64	78.31	70.95	0.3722	0.1721	0.3471	**0.7622**	**0.4892**	**0.1197**
	12	MLP	77.61	80.44	77.61	78.57	73.01	0.3353	0.1515	0.3456	0.7751	0.5168	0.1028
MRI	13	RF	47.53	46.27	47.53	46.30	45.96	0.3535	0.1639	0.4257	0.5168	0.3004	0.2662
	14	SVM	48.06	45.92	48.06	45.78	46.81	0.3506	0.1666	0.4152	0.5938	0.3565	0.2403
	15	**XGB (13)**	**54.13**	**54.08**	**54.13**	**53.93**	**53.27**	0.2363	0.1088	0.4530	0.5141	0.2964	0.2678
	16	KNN	40.78	44.98	40.78	39.81	41.29	0.2528	0.1150	0.4721	0.4896	0.2702	0.2799
	17	DT	40.78	41.48	40.78	41.07	36.02	0.1479	0.0660	**0.4934**	**0.3349**	**0.1729**	**0.3277**
	18	MLP	48.81	48.36	48.81	48.49	46.93	0.1959	0.0887	0.4602	0.3950	0.2138	0.3091
Static	19	RF	53.89	53.45	53.89	53.05	49.09	0.2680	0.1259	0.4208	0.5978	0.3518	0.2407
	20	SVM	52.67	55.44	52.67	52.73	49.57	0.2797	0.1335	0.4354	0.6172	0.3787	0.2294
	21	**XGB (12)**	**58.25**	**57.28**	**58.25**	**57.31**	**53.28**	0.2941	0.1388	0.3929	0.6427	0.3798	0.2302
	22	KNN	43.45	47.74	43.45	43.08	41.96	0.3608	0.1708	0.4206	0.5835	0.3480	0.2447
	23	DT	48.53	49.92	48.53	49.05	45.71	**0.1307**	**0.0605**	0.4616	**0.4651**	**0.2558**	**0.2892**
	24	MLP	52.86	54.26	52.86	53.11	49.69	0.2635	0.1229	0.4312	0.5558	0.3234	0.2553
CSs + NSB	25	RF	82.61	83.86	82.61	83.08	77.61	0.5309	0.2572	0.3046	0.5877	0.3231	0.1917
	26	SVM	65.78	71.10	65.78	67.89	61.86	0.4436	0.2522	0.3311	0.4681	0.2381	0.2484
	27	**XGB (2)**	**85.68**	**86.37**	**85.68**	**86.00**	**80.17**	0.5354	0.2613	0.3026	0.5961	0.3242	0.1901
	28	KNN	54.37	60.71	54.37	56.58	53.14	0.3933	0.2357	0.3687	**0.4153**	**0.2116**	**0.3050**
	29	DT	80.07	82.49	80.07	81.12	74.50	0.4784	0.2219	0.3218	0.5392	0.2719	0.2120
	30	MLP	69.17	76.75	69.17	67.73	69.27	0.5121	0.2576	0.3106	0.6003	0.3335	0.1942
CSs + MRI	31	RF	79.12	80.72	79.12	79.59	74.93	0.5086	0.2461	0.3111	0.5923	0.3308	0.2213
	32	SVM	56.07	58.74	56.07	57.20	51.72	0.4012	0.2332	0.3658	0.4633	0.2403	0.2876
	33	**XGB (5)**	**81.55**	**81.96**	**81.55**	**81.62**	**76.27**	0.5422	0.2685	0.3035	0.6056	0.3491	0.2148
	34	KNN	50.00	56.30	50.00	52.03	48.80	0.3730	0.2262	0.3863	**0.3935**	**0.2053**	**0.3224**
	35	DT	74.30	75.21	74.30	74.67	67.41	0.4643	0.2332	0.3301	0.5582	0.3134	0.2310
	36	MLP	70.09	77.44	70.09	69.98	69.38	0.5177	0.2624	0.3287	0.5300	0.2801	0.2512
CSs + Static	37	RF	79.63	80.83	79.63	80.02	74.48	0.5150	0.2575	0.3080	0.5479	0.3092	0.2379
	38	SVM	52.62	58.39	52.62	54.91	49.41	0.3878	0.2300	0.3766	0.3898	0.2090	0.3115
	39	**XGB (4)**	**82.52**	**83.44**	**82.52**	**82.91**	**77.16**	0.4958	0.2501	0.3090	0.5436	0.3157	0.2362
	40	KNN	48.06	54.35	48.06	50.08	47.38	0.3340	0.2019	0.4040	**0.3423**	**0.1849**	**0.3382**
	41	DT	77.84	78.94	77.84	78.26	72.10	0.4687	0.2342	0.3238	0.5129	0.2891	0.2460
	42	MLP	66.16	73.10	66.16	63.84	65.32	0.1896	0.0898	0.4174	0.3915	0.1833	0.3050
NSB + MRI	43	RF	75.43	76.25	75.43	75.55	72.05	0.4525	0.2249	0.3316	0.4685	0.2562	0.2557
	44	SVM	63.83	67.00	63.83	65.09	59.99	0.4154	0.2392	0.3414	0.4130	0.2275	0.2743
	45	**XGB (9)**	**78.88**	**80.56**	**78.88**	**79.53**	**76.21**	0.4914	0.2439	0.3145	0.5767	0.3145	0.2298
	46	KNN	51.70	58.04	51.7	53.79	50.09	0.4035	0.2435	0.3732	0.3540	0.1972	0.3204
	47	DT	73.94	75.39	73.94	74.44	70.54	0.4324	0.2098	0.3383	0.4613	0.2543	0.2573
	48	MLP	65.60	70.35	65.60	63.91	65.20	0.1908	0.0893	0.4271	**0.3326**	**0.1590**	**0.3212**
NSB + Static	49	RF	77.26	77.78	77.26	77.39	72.53	0.4733	0.2353	0.3209	0.4679	0.2740	0.2508
	50	SVM	61.17	67.77	61.17	63.66	57.25	0.3822	0.2262	0.3540	0.2856	0.1586	0.3107
	51	**XGB (7)**	**80.58**	**80.74**	**80.58**	**80.66**	**75.43**	0.4798	0.2379	0.3149	0.4642	0.2601	0.2549
	52	KNN	49.76	55.43	49.76	51.68	48.25	0.3702	0.2242	0.3880	0.2764	0.1565	0.3439
	53	DT	75.22	75.16	75.22	75.17	68.31	0.4181	0.2033	0.3359	0.4206	0.2399	0.2646
	54	MLP	64.75	69.29	64.75	62.67	63.54	0.3456	0.1681	0.4020	**0.2737**	**0.1236**	**0.3471**
MRI + Static	55	RF	54.97	54.14	54.97	53.84	52.85	0.3925	0.1873	0.3965	0.4851	0.2806	0.2767
	56	SVM	48.79	50.92	48.79	49.70	43.69	0.3797	0.2134	0.3974	0.4062	0.2272	0.3058
	57	**XGB (11)**	**65.05**	**64.07**	**65.05**	**64.33**	**58.58**	0.4211	0.2038	0.3589	0.5690	0.3245	0.2581
	58	KNN	44.66	49.78	44.66	46.18	43.80	0.3409	0.2037	0.4143	**0.3624**	**0.2021**	**0.3228**
	59	DT	50.40	51.17	50.40	50.72	47.20	0.2229	0.1014	0.4398	0.3825	0.2076	0.3131
	60	MLP	50.45	54.80	50.45	46.82	51.04	0.4581	0.2240	0.3665	0.4750	0.2713	0.2807
CSs + NSB + MRI	61	RF	82.62	84.19	82.62	83.13	78.77	0.5472	0.2763	0.3016	0.5221	0.2762	0.2407
	62	SVM	62.86	66.33	62.86	64.16	59.43	0.3835	0.2286	0.3482	**0.3381**	**0.1815**	**0.2933**
	63	**XGB (3)**	**83.74**	**85.18**	**83.74**	**84.32**	**78.76**	0.5529	0.2725	0.3001	0.5389	0.2801	0.2391
	64	KNN	50.00	56.36	50.00	52.14	48.85	0.3710	0.2300	0.3843	0.3629	0.1914	0.3305
	65	DT	78.18	80.36	78.18	79.06	74.00	0.4649	0.2190	0.3260	0.4346	0.2185	0.2625
	66	MLP	65.39	73.41	65.39	63.28	64.67	0.3980	0.1957	0.3866	0.3614	0.1609	0.3289
CSs + NSB + Static	67	RF	83.22	84.34	83.22	83.63	78.33	0.5231	0.2593	0.3055	0.5392	0.2912	0.2180
	68	SVM	62.38	67.35	62.38	64.29	59.22	0.4419	0.2584	0.3370	0.4265	0.2193	0.2731
	69	**XGB (1)**	**85.92**	**86.95**	**85.92**	**86.37**	**81.11**	0.5322	0.2587	0.3028	0.6016	0.3088	0.2108
	70	KNN	49.51	55.43	49.51	51.44	48.68	0.3578	0.2197	0.3909	**0.3271**	**0.1738**	**0.3410**
	71	DT	82.23	83.55	82.23	82.82	75.91	0.4443	0.2076	0.3248	0.4785	0.2387	0.2358
	72	MLP	66.56	73.02	66.56	63.08	65.78	0.5327	0.2711	0.3189	0.4874	0.2432	0.2528
NSB + MRI + Static	73	RF	76.21	76.95	76.21	76.35	72.79	0.4611	0.2302	0.3248	0.4603	0.2521	0.2544

Table 4 (*continued*)

Modality	#	Model	Performance (%)					Overall Diversity			Diversity of each data set		
			Acc	Pre	Rec	F1	BA	Q.	Corr.	Dis.	Q.	Corr.	Dis.
	74	SVM	60.92	64.70	60.92	62.24	58.10	0.3792	0.2284	0.3535	0.3625	0.1962	0.2933
	75	**XGB (8)**	**79.61**	**80.04**	**79.61**	**79.80**	**73.31**	0.4999	0.2538	0.3107	0.5257	0.2857	0.2399
	76	KNN	47.82	53.92	47.82	49.83	46.96	0.3499	0.2160	0.3978	**0.3072**	**0.1707**	**0.3443**
	77	DT	72.09	72.75	72.09	72.21	67.72	0.4322	0.2106	0.3397	0.4720	0.2497	0.2577
	78	MLP	64.18	68.38	64.18	60.38	63.25	0.5440	0.2778	0.3258	0.4571	0.2352	0.2731

Abbreviations: Acc, accuracy; Pre, precision; Rec, recall; F1, F1-score; Q., Q-statistic; Corr., correlation coefficient; Dis., disagreement.

used a greedy method to select the optimal combination. We started by fusing two base models. To build a voting ensemble, we used soft voting because, in most cases, it achieves better results [77]. In addition, every experiment was repeated 30 times to get more stable results. Table 5 shows the testing results of different ensembles. We implemented both voting and stacking ensembles, we start by analyzing the results of the voting models. Note that the implemented customized ensembles are based on the accuracy and diversity of the base classifiers.

First, we analyzed the decision fusion of base classifiers trained using single modalities (i.e., models 1 to 24 in Table 4). These models were trained independently using CSs, NSB, MRI, or Static data. The two most accurate base models were 7 (CSs RF) and 9 (CSs XGB). A voting classifier based on these two models improved their individual performance (Accuracy = 85.57 %, Precision = 86.17 %, Recall = 85.57 %, F1-score = 85.71 %, and BA = 81.02 %). However, replacing model 7 by model 3 (NSB XGB), which is less accurate but more diverse, improves the performance of this voting ensemble (Accuracy = 87.38 %, Precision = 87.67 %, Recall = 87.38 %, F1-score = 87.43 %, and BA = 82.61 %). Adding model 21 to the 3–9 ensemble achieved better performance than the previous early fusion scheme (Accuracy = 87.86 %, Precision = 87.74 %, Recall = 87.86 %, F1-score = 87.73 %, and BA = 82.98 %). We note that replacing model 21 with the most diverse model (model 23) degrades the performance of the previous 3–9-21 ensemble. This is expected because the three models (3, 9, and 21) are trained using different feature sets (i.e., CSs, NSB, and Static). Therefore, they are already diverse and replacing a more accurate model with a less accurate one would surely harm the results. In other words, the DT model (23) did not add to the diversity of the ensemble. Other combinations have been checked, but no ensemble beats the 3–9-21 model.

Second, we analyze the decision fusion of base classifiers trained using two fused modalities (i.e., we pay attention here to models 25 to 60 in Table 4). The XGB classifiers achieved the best results, especially model 27 (Accuracy = 85.68 %, Precision = 86.37 %, Recall = 85.68 %, F1-score = 86.00 %, and BA = 80.17 %). The fusion

Table 5
Customized ensemble selection.

Datasets	Fusion strategy	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	BA (%)
NSB, CSs	Vote: 2 most accurate	3, 9	87.38	87.67	87.38	87.43	82.61
CSs	Vote: 2 most accurate	7, 9	85.57	86.17	85.57	85.71	81.02
NSB, CSs, static	Vote: 3 most accurate and diverse	3, 9, 21	**87.86**	**87.74**	**87.86**	**87.73**	**82.98**
NSB, CSs, MRI, Static	Vote: 4 most accurate	3, 9, 15, 21	86.89	86.69	86.89	86.67	81.42
NSB, CSs, MRI, Static	Vote: 4 most diverse	5, 11, 17, 23	75.85	75.74	75.85	74.75	70.78
NSB, CSs, Static	Vote: 2 most accurate + 1 most diverse	3, 9, 23	84.81	84.80	84.81	84.67	80.63
[CSs, NSB] + [CSs, Static]	Vote: 2 most accurate	27, 39	86.89	87.56	86.89	87.10	83.72
[CSs, NSB] + [NSB, Static]	Vote: 1 most accurate + 1 diverse	27, 51	**88.35**	**88.58**	**88.35**	**88.46**	**83.55**
[CSs, NSB] + [NSB, MRI]	Vote: 1 most accurate + 1 more diverse	27, 45	86.65	87.17	86.65	86.89	81.69
[CSs, NSB] + [MRI, Static]	Vote: 1 most accurate + 1 more diverse	27, 57	86.89	86.88	86.89	86.86	82.00
[CSs, NSB] + [CSs, NSB]	Vote: 1 most accurate + 1 accurate	27, 25	87.51	88.41	87.51	87.86	83.49
[CSs, NSB] + [CSs, NSB]	Vote: 1 most accurate + 1 more diverse	27, 28	87.14	90.23	87.14	87.90	86.93
[CSs, NSB] + [CSs, Static]	Vote: 1 most accurate + 1 more diverse	27, 38	86.65	88.19	86.65	87.16	83.85
[CSs, NSB] + [CSs, NSB] + [CSs, Static]	Vote: 3 most accurate	27, 25, 39	87.09	87.58	87.09	87.25	83.11
[CSs, NSB] + [CSs, Static] + [NSB, Static]	Vote: 2 most accurate + 1 diverse	27, 39, 51	**88.59**	**88.80**	**88.59**	**88.68**	**83.84**
[CSs, NSB, Static] + [CSs, NSB, MRI]	Vote: 2 most accurate	63, 69	87.14	88.08	87.14	87.51	83.08
[CSs, NSB, Static] + [CSs, NSB, Static]	Vote: 1 most accurate + 1 diverse	67, 69	**88.09**	**88.90**	**88.09**	**88.41**	**84.32**
NSB, CSs	Stack: 2 most accurate and diverse	3, 9	86.65	86.43	86.65	86.51	80.36
CSs	Stack: 2 most accurate	7, 9	85.78	85.94	85.78	85.83	80.00
CSs + Static	Stack: 1 most accurate + 1 most diverse	9, 23	50.72	52.72	50.72	51.54	47.15
NSB, CSs, static	Stack: 3 most accurate and diverse	3, 9, 21	**88.11**	**87.55**	**88.11**	**87.78**	**81.46**
CSs	Stack: 3 most accurate	7, 9, 11	77.02	77.17	77.02	76.95	69.83
NSB, CSs, MRI, Static	Stack: 4 most accurate	7, 9, 11, 12	81.66	81.95	81.66	81.70	75.59
NSB, CSs, MRI, Static	Stack: 4 most accurate and diverse	3, 9, 15, 21	87.62	86.81	87.62	87.10	80.25
NSB, CSs, MRI, Static	Stack: 4 most diverse	5, 11, 17, 23	75.39	76.18	75.39	75.67	69.14
NSB, CSs, Static	Stack: 2 most accurate + 1 most diverse	3, 9, 23	73.39	73.14	73.39	72.53	68.52
[CSs, NSB] + [CSs, Static]	Stack: 2 most accurate	27, 39	87.14	87.41	87.14	87.24	82.64
[CSs, NSB] + [NSB, Static]	Stack: 1 most accurate + 1 diverse	27, 51	**88.35**	**88.41**	**88.35**	**88.38**	**83.55**
[CSs, NSB] + [NSB, MRI]	Stack: 1 most accurate + 1 more diverse	27, 45	86.65	86.86	86.65	86.73	80.85
[CSs, NSB] + [CSs, NSB]	Stack: 1 most accurate + 1 accurate	27, 25	87.56	88.11	87.56	87.79	82.80
[CSs, NSB] + [MRI, Static]	Stack: 1 most accurate + 1 more diverse	27, 57	87.86	88.17	87.86	87.98	82.70
[CSs, NSB] + [CSs, NSB]	Stack: 1 most accurate + 1 more diverse	27, 28	88.35	89.14	88.35	88.67	84.85
[CSs, NSB] + [CSs, Static]	Stack: 1 most accurate + 1 more diverse	27, 38	87.86	88.43	87.86	88.10	83.87
[CSs, NSB] + [CSs, NSB] + [CSs, Static]	Stack: 3 most accurate	27, 25, 39	87.82	88.17	87.82	87.96	83.35
[CSs, NSB] + [CSs, Static] + [NSB, Static]	Stack: 2 most accurate + 1 diverse	27, 39, 51	**88.83**	**88.72**	**88.83**	**88.77**	**83.59**
[CSs, NSB, Static] + [CSs, NSB, MRI]	Stack: 2 most accurate	63, 69	87.14	87.98	87.14	87.48	83.06
[CSs, NSB, Static] + [CSs, NSB, Static]	Stack: 1 most accurate + 1 diverse	67, 69	87.86	88.46	87.86	88.12	83.51
2[CSs, NSB, Static] + [CSs, NSB, MRI]	Stack: 3 most accurate	69, 63, 67	87.64	88.21	87.64	87.88	83.26

of this model (27) and the second-most accurate model (39) improves the performance (Accuracy = 86.89 %, Precision = 87.56 %, Recall = 86.89 %, F1-score = 87.10 %, and BA = 83.72 %). However, replacing the most accurate model 27 with a less accurate one that is more diverse, like model 51, further improves the performance (Accuracy = 88.35 %, Precision = 88.58 %, Recall = 88.35 %, F1-score = 88.46 %, and BA = 83.55 %). In addition, the NSB data are utilized by both models 27 and 51. The role of diversity is also confirmed by replacing model 51 with a very weak model (45), and the resulting ensemble achieved even better performance than model 27 alone (Accuracy = 86.65 %, Precision = 87.17 %, Recall = 86.65 %, F1-score = 86.89 %, BA = 81.69 %). As we increase diversity, the resulting vote classifier achieves better results, as shown in Table 5. For example, in model 28, the KNN based classifier only has an accuracy of 54.37 % but it has a high diversity level, as shown in Table 4. Combining model 28 with model 27 achieved superior results (Accuracy = 87.14 %, Precision = 90.23 %, Recall = 87.14 %, F1-score = 87.9 %, BA = 86.93 %), especially for Precision and BA. All other combinations were checked. Although some of them enhanced the performance of participating base classifiers, no other ensemble beats the 27–51 model. We noticed that adding more base classifiers to the two-base model scheme did not enhance performance.

Third, we analyzed the decision fusion of base classifiers trained using three fused modalities. As expected, the integrated CSs, NSB, and Static dataset achieved the best results with the XGB classifier (Accuracy = 85.92 %, Precision = 86.95 %, Recall = 85.92 %, F1-score = 86.37 %, and BA = 81.11 %). Note that we achieved better results using the late fusion of the same modalities, as explained before, with models 27–51. We checked the fusion of this model and other accurate or diverse models. The diversity has shown a better effect to improve the performance, as in 67–69 models. All other combinations did not achieve better results.

Next, we study the results of stacking models. Based on the base modalities, the stacking of the most accurate three models, 3, 9, and 21, achieved better results than the equivalent voting model (Accuracy = 88.11 %, Precision = 87.55 %, Recall = 88.11 %, F1-score = 87.78 %, and BA = 81.46 %). Note that combining more accurate base models (7, 9, and 11) generates a less accurate ensemble because these models have lower diversity than 3, 9, and 21. Like voting, increasing the number of base models increases the computational ensemble time, but it does not improve the results of stacking models. We checked on the role of diversity in enhancing the ensemble performance. The best stacking model is the result of fusing the three models 27, 39, and 51, which are based on [CSs, NSB], [CSs, Static], and [NSB, Static], respectively (Accuracy = 88.8 3 %, Precision = 88.72 %, Recall = 88.83 %, F1-score = 88.77 %, and BA = 83.59 %). It is worth noting that these results are better than all voting models.

These results indicate that the customized stacking model based on the decision fusion of heterogeneous data achieved the most accurate results. The XGB classifiers were used to train three base models. Then, a stacking ensemble model was used to fuse the decisions of these three base models using LR as the meta learner. It can be noticed that CSs and NSB are early fused and used by one XGB classifier. In addition, the Static data are fused with both CSs and NSB in two different datasets and used to train the other two XGB classifiers. As a result, each modality is used twice.

3.2.3. Optimizing selected model

The literature asserts that the performance of base learners must be optimized before using ensemble models [78]. This optimization can be achieved either using the base learners in meta classifiers such as voting or stacking or by hyperparameter optimization [79], but not necessarily using both. In the previous sections, we considered the default settings of the XGB classifiers. In this experiment, we check the effect of optimizing the performance of the base learners on the overall ensemble results. Based on the results of the previous experiments, we optimized an XGB base classifier for each dataset of [CSs, NSB], [CSs, Static], and [NSB, Static]. A randomized search with a stratified 10-CV strategy was used to find the best hyperparameters for every XGB classifier. The classifiers were fitted with the MDS and stratified with 10-CV. Each experiment was repeated 30 times, the average performance of each model was measured using the unseen MTS test set. Next, we combined these optimized models to train the stacking ensemble model. Optimization results are presented in Table 6 and exhaustive lists of the hyperparameters are in Table S2 of Supplementary File 2. As can be seen, optimizing the hyperparameters slightly improves the performance of the base classifiers. This enhanced performance of the base classifiers has no big effect on the overall performance of the stacking model. This is probably because the diversity of the base classifiers has not been improved.

3.2.4. Testing optimized model on 3-class and 2-class problems

The previously trained models were used to solve a 4-class classification problem (i.e., CN vs sMCI vs pMCI vs AD). This is a problematic classification task [6,9]. The main challenge is to identify the pMCI patients. As discussed in Section 3.1 Statistical Analysis, the pMCI cases have similar properties to the AD class at M48. In fact, they will be AD patients at M84, but from BL to M18, they are MCI patients with worse conditions, see Figs. 4, 5, 6, and 7. As a result, these cases induce noise to the classification algorithms because they are still MCI, but classifiers have to discriminate them as a different category.

In this experiment, we test our proposal regarding the 3-class classification problem (i.e., CN vs MCI vs AD). We consider pMCI and AD to be in the same category, sMCI as just stable MCI, and CN as it is. In addition, we test our approach in a 2-class classification problem (i.e., NOT AD vs AD). In practice, we check all possible combinations between two classes, namely CN vs AD (AD is the positive class), CN vs pMCI (pMCI is the positive class), CN vs sMCI (sMCI is the positive class), AD vs sMCI (AD is the positive class), AD vs pMCI (AD is the positive class), and sMCI vs pMCI (pMCI is the positive class). Notice that we set the class nearer to AD as the positive class. In addition, we check the binary classification

Table 6
Optimized performance of base XGB models and stacking model.

Model	Hyperparameters	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	BA (%)
[CSs, NSB] model	Default	85.68	86.37	85.68	86.00	80.17
	Optimized	87.90	88.52	87.90	88.13	84.04
[CSs, Static] model	Default	82.52	83.44	82.52	82.91	77.16
	Optimized	83.21	83.21	83.21	83.12	77.59
[NSB, Static] model	Default	80.58	80.74	80.58	80.66	75.43
	Optimized	82.22	83.15	82.22	82.59	77.81
Stacking model	Default	88.83	88.72	88.83	88.77	83.59
	Optimized	**89.15**	**89.11**	**89.16**	**89.07**	**84.44**

problem when combining both CN and sMCI as class 0 and pMCI and AD as class 1.

As reported in Table 7, both base and ensemble classifiers achieve higher performance compared to the 4-class problem. It is worth noting that the used modalities and base classifiers in this experiment are based on the results of the previous experiment. For the 3-class problem, the stacking classifier achieved an accuracy of 96.43 %, precision of 96.47 %, recall of 96.43 %, F1-score of 96.44 %, and BA of 96.38 %. Compared to the base learners, it can be seen that the stacking model achieves better performance for all classification tasks. It is worth noting that the best results achieved by the base learners are due to careful data preprocessing, feature engineering, and hyperparameter optimization steps. The results of the sMCI vs pMCI and [CN/sMCI] vs [pMCI/AD] binary classification tasks assert that pMCI cases are challenging to identify. However, the proposed stacking ensemble model has the highest performance in all 3-class and 2-class classification tasks.

3.2.5. Comparison with the state-of-the-art from previous literature

This section presents a comparison between the best proposed competitive ensemble model in previous literature for AD progression detection, especially ensemble models. Please note that we are considering accuracy as the common metric that is given in all related works for the sake of fairness. Unfortunately, some authors do not publish the data they have used, and it is not possible to reproduce their experiments here. We just trust what was published by other authors to make an illustrative comparison. The proposed model is based on the stacking technique, which is like a voting ensemble except that the decisions of the base classifiers are used by another meta classifier to determine the weight of each learner. We utilize the early fusion of data by testing all possible combinations of the four modalities and selecting the best results. Level-0 in the proposed stacking ensemble has three XGB models as base classifiers with inputs of [CSs, NSB], [CSs, Static], and [NSB, Static]. Then, the decisions of these models are fused and trained using an LR algorithm as the meta classifier for Level-1.

Furthermore, the proposed stacking model achieved the best performance because it is based on a carefully selected list of features extracted from different modalities of time series data. Then, our proposed model integrates the capabilities of voting and boosting techniques using XGB as the base learner and stacking as the decision fuser.

As previously noted, this model achieved optimal performance without considering MRI neuroimaging features. This means that the model is expected to be less expensive because it is based on a less expensive list of features. As a result, the model is applicable in a wider variety of medical environments and is especially suitable for use in developing countries. Table 8 shows a comparison of different AD ensemble models. It can be seen that the majority of the studies concentrated on MRI data analysis, yet they achieved worse results than our model [5,6,9,23,24,81,83,85,87,88,94,93,97–98]. The data fusion of multiple modalities has not been tested in the literature. In addition, most studies concentrated on building ensemble models to solve binary classification problems such as CN vs MCI [90,95], CN vs AD [38,102,103,100,97,93], sMCI vs pMCI [99], or AD vs MCI [98]. Multiclass classification problems are either converted into multiple binary classification problems [42,94,101], or implemented as an *n*-class classification problem, where *n* could be three [92] or four [5,6,9,24,44,91]. For *n*-class classification tasks, the literature has reported lower accuracy than seen here. For example, Yao et al. [6] used only the baseline MRI data and three demographics to build a hierarchical ensemble classifier for a 4-class prediction task. The model achieved an accuracy of 54.38 % based on XGB and SVM classifiers.

As that study was based on a small number of modalities and no time-series data, it achieved only poor performance. Donnelly-

Table 7
The 3-class and 2-class results.

Model	No. of classes	Performance (%)				
		Accuracy	Precision	Recall	F1-score	BA
[CSs, NSB] model	CN vs AD	98.56	98.56	98.56	98.56	98.51
	CN vs pMCI	97.62	97.62	97.62	97.60	96.03
	CN vs sMCI	98.11	98.13	98.11	98.11	98.03
	AD vs sMCI	96.11	96.12	96.11	96.10	95.80
	AD vs pMCI	93.80	93.76	93.80	93.73	91.47
	sMCI vs pMCI	82.55	82.26	82.55	82.39	74.33
	[CN/sMCI] vs [pMCI/AD]	89.26	89.62	89.26	89.36	89.17
	CN vs MCI vs AD	95.76	95.81	95.76	95.76	95.68
[CSs, Static] model	CN vs AD	99.56	99.56	99.56	99.56	99.51
	CN vs pMCI	97.62	97.62	97.62	97.60	96.03
	CN vs sMCI	94.45	94.63	94.45	94.44	94.23
	AD vs sMCI	93.85	93.86	93.85	93.86	93.73
	AD vs pMCI	83.70	84.16	83.70	83.88	81.45
	sMCI vs pMCI	81.83	81.86	81.83	81.84	74.28
	[CN/sMCI] vs [pMCI/AD]	89.47	89.49	89.47	89.48	88.53
	CN vs MCI vs AD	93.67	93.66	93.67	93.65	93.66
[NSB, Static] model	CN vs AD	98.67	98.67	98.67	98.67	98.61
	CN vs pMCI	96.45	96.54	96.45	96.48	96.04
	CN vs sMCI	92.85	93.11	92.85	92.82	92.58
	AD vs sMCI	90.05	89.91	90.05	89.91	86.67
	AD vs pMCI	89.91	89.77	89.91	89.77	86.57
	sMCI vs pMCI	82.55	82.05	82.55	82.26	73.66
	[CN/sMCI] vs [pMCI/AD]	88.99	89.26	88.99	89.07	88.70
	CN vs MCI vs AD	89.97	90.16	89.97	90.02	90.23
Stacking model	**CN vs AD**	**99.56**	**99.56**	**99.56**	**99.56**	**99.51**
	CN vs pMCI	**98.08**	**98.07**	**98.08**	**98.07**	**96.94**
	CN vs sMCI	**98.12**	**98.15**	**98.12**	**98.12**	**98.05**
	AD vs sMCI	**96.14**	**96.15**	**96.14**	**96.13**	**95.81**
	AD vs pMCI	**94.56**	**94.57**	**94.56**	**94.56**	**93.48**
	sMCI vs pMCI	**85.13**	**84.78**	**85.13**	**84.92**	**77.56**
	[CN/sMCI] vs [pMCI/AD]	**90.06**	**90.17**	**90.06**	**90.10**	**89.47**
	CN vs MCI vs AD	**96.43**	**96.47**	**96.43**	**96.44**	**96.38**

Table 8
A comparison with other ensemble studies in the literature.

Reference	Population	Modalities	Classes	Time series	ML method	Accuracy	Task
Proposed ensemble	**ADNI: 1371 subjects**	**CSs + NSB + Static**	**4**	**YES**	**Stacking with XGB**	**89.15 %**	**Progression detection**
Tanveer et al. [80] (2021)	ADNI: 813 subjects	MRI	2	NO	Voting with three deep neural networks	85.27 %	CN/AD diagnosis
Liu et al. [81] (2021)	ADNI: 2,228 subjects	MRI + Age + Gender + MMSE	2	NO	Voting with nine graph CNN	83.50 %	Progression detection
Giovannetti et al. [82] (2021)	Hospital Universitario San Carlos (Spain): 87 subjects	Magnetoencephalography Recordings + MRI	3	YES	Voting with AlexNet + {SVM + LDA}	87.00 %	Progression detection
BI et al. [83] (2021)	ADNI: 116 subjects	fMRI	2	NO	Clustering-evolutionary weighted SVMs	CN/MCI (83.47 %), sMCI/pMCI (84.30 %)	Progression detection
Niyas and	Thiyagarajan [84] (2021)	ADNI: 1737 subjects	MRI, PET, CSF,		cognitive tests, age, sex. and education	3	NO
Dynamic ensemble	84 %	AD diagnosis					
Ruiz et al. [85] (2021)	ADNI: 480 subjects	MRI	4	NO	Voting with three 3D Densely Connected CNNs	83.33 %	Progression detection
An et al. [86] (2020)	NACC: 23,165 samples	MH, HIS, CVD, UPDRS, NPIQ, GDS, FAQ	2	NO	Stacking with neural network as meta classifier and sparse autoencoders as base classifiers	87 %	AD diagnosis
Pan et al. [87] (2020)	ADNI: 509 subjects	MRI	2	NO	Multilayer voting with CNN (3D-SENet) base classifiers	CN/AD (84 %), CN/pMCI (79 %), sMCI/pMCI (62 %)	Progression detection
Choi et al. [88] (2020)	ADNI: 815 subjects	MRI	3	NO	Voting with CNN (VGG-16, GoogLeNet, and AlexNet)	AD/MCI/NC (93.84 %)	AD diagnosis
SYED et al. [89], (2020)	Figshare: 333 subjects	CSF protein biomarkers, age, gender, amyloid proteins, tau, ptau, Aβ42	2	NO	Weighted average with two base learners of LR and Linear SVM	95.52 %	CN/MCI diagnosis
Yao et al. [6] (2018)	ADNI: 400 subjects	MRI + Age + Gender + MMSE	4	NO	Hierarchical ensemble of 5 binary classifiers by XGB and SVM	54.38 %	Progression detection
Nanni et al. [24] (2018)	ADNI: 400 subjects	MRI + Age + MMSE	4	NO	Voting with sum rule of 6 classifiers	52.92 %	Progression detection
Sorensen et al. [44] (2018)	ADNI: 400 subjects	MRI + Age + Gender + MMSE	4	NO	Bagging with SVM	59.10 %	Progression detection
Dimitriadis and Liparas [9] (2018)	ADNI: 400 subjects	MRI	4	NO	Voting with RF	61.90 %	Progression detection
Qiu et al. [90] (2018)	NACC: 386 subjects	MRI, MMSE, logical memory (LM)	2	NO	Voting of CNN (VGG-11) for MRI with MLP of MMSE and LM	90.90 %	MCI/CN diagnosis
Ramírez et al. [5] (2018)	ADNI: 740 subjects	MRI	4	NO	Bagging with random forest one vs rest classifiers	56.25 %	Progression detection
Jin and Deng [91] (2018)	ADNI: 400 subjects	MRI, age, gender, MMSE	4	NO	Gradient boosting tree	56.25 %	Progression detection
Ebadi et al. [42] (2017)	ADNI: 45 subjects	Diffusion Tensor Imaging (DTI)	2	NO	Voting with LR, RF, SVM, KNN, and Gaussian NB	AD/MCI (83.3 %), AD/CN (80 %), MCI/CN (70 %)	AD diagnosis
Moore et al. [92] (2017)	ADNI: 1737 subjects	10 features from MRI, CSs, and CSF	3	YES	Random forest	73.00 %	AD/MCI/CN diagnosis
Armananzas et al. [93] (2017)	ADRC: 41	fMRI	2	NO	Random linear oracle with SVM	95.00 %	AD/CN diagnosis
Nanni et al. [94] (2016)	ADNI: 509 subjects	MRI	2	NO	Voting with weight sum rule and SVM	AD/CN (93.3 %), pMCI/CN (88.9 %), pMCI/sMCI (68.6 %)	AD diagnosis
Sivapriya et al. [95] (2015)	ADNI: 750	NSB + FDG PET	2	NO	Voting with C4.5 for feature selection then C4.5 for classification	97.60 %	CN/MCI diagnosis
Farhan et al. [38] (2014)	OASIS: 85 subjects	MRI	2	NO	Voting with on SVM, RF, and MLP	93.75 %	AD/CN diagnosis
Gray et al. [96] (2013)	ADNI: 147 subjects	MRI + FDG PET + CSF + genetics	2	NO	Random forest and early fusion	AD/CN (89 %), MCI/CN (75 %)	AD diagnosis

Table 8 (*continued*)

Reference	Population	Modalities	Classes	Time series	ML method	Accuracy	Task
Simpson et al. [97] (2013)	ADNI: 311	MRI	2	NO	Bagging with SVM	85.90 %	AD/CN diagnosis
Liu et al. [98] (2012)	ADNI: 652 subjects	MRI	2	NO	Voting based on sparse representation-based classifier (SRC)	90.80/87.85 % for AD/MCI	AD/MCI diagnosis
Chen et al. [99] (2012)	ADNI: 26	MRI	2	NO	Boosting with Bayesian-network representation	81.00 %	sMCI/pMCI diagnosis
Polikar et al. [100] (2010)	ADNI: 300 subjects	EEG, MRI, PET	2	NO	Voting with 15 MLP classifiers	85.55 %	AD/CN diagnosis
Silveira and Marques [101] (2010)	ADNI: 268	PET	2	NO	Boosting	AD/CN (90.97 %), MCI/CN (79.63 % %), MCI/AD (70 %)	AD diagnosis
Hinrichs et al. [102] (2009)	ADNI: 183 subjects	MRI + FDG PET	2	NO	Boosting with linear program (LP)	82.00 %	AD/CN diagnosis
Polikar et al. [19] (2008)	MDCUP: 52 subjects	EEG	2	NO	Voting for ensemble-of-ensembles with Learn^{++}	79.20 %	Progression detection
Gandhi et al. [103] (2006)	MDCUP: 80 subjects	EEG	2	NO	Stacked generalization	76.40 %	AD/CN diagnosis

Abbreviations: MDCUP, Memory Disorders Clinic of the University of Pennsylvania; NB, Naïve Bayes; ADRC, Washington University Alzheimer's Disease Research Center; NACC, National Alzheimer's Coordinating Center; medical history (MH), Hachinski ischemic score (HIS), cerebrovascular disease (CVD), Unified Parkinson's Disease Rating Scale (UPDRS), Neuropsychiatric Inventory Questionnaire (NPIQ), Geriatric Depression Scale (GDS), Functional Activities Questionnaire (FAQ).

Kehoe et al. [22] asserted that the maximum accuracy achieved using MRI features is lower than the accuracy achieved using the MMSE feature alone. This means that other modalities like CSs are critical. Nanni et al. [24] used baseline MRI data and two other features, age, and MMSE. They built a voting classifier based on the sum rule of six classifiers, and they achieved an accuracy of 52.92 %. Furthermore, most studies have not considered the temporal dimension of AD data. As we asserted before, the AD data are multimodal and time series in nature. As a result, neglecting the longitudinal dimension in AD data analysis must negatively affect the system performance. Moore et al. [92] used time series data to build an AD detection model but for a simpler task than the progression detection task we tackled. They used the RF algorithm and achieved an accuracy of 73 % for the CN vs MCI vs AD task. As far as we know, the proposed model is the most accurate and comprehensive system in the literature so far. The most relevant properties of our approach are: (1) The size of our dataset is significantly larger than most studies in Table 8; (2) we have considered the multimodality and time-series nature of the data; (3) we carefully selected the most discriminating features from the used modalities; (4) different ensemble techniques have been tested, and the stacking ensemble model with XGB achieved the best performance; (5) the selected model achieved a balance between accuracy and diversity among the selected base learners; and, finally, (6) the best results were achieved without considering MRI data, making our model cheaper and medically more widely applicable.

However, our model has some limitations that will be handled in future work. For example, the current study concentrates mainly on performance. However, in real medical environments, domain experts insist on interpretable results in addition to accurate models. We will extend this study to evaluate and enhance our model's interpretable features. In addition, we did not consider an optimization technique for selecting the most relevant base classifiers. This is mainly because the number of base classifiers is few. As a result, an exhaustive search was considered a feasible solution for classifier selection. In a future study, we will consider the fusion of other modalities (e.g., medications, comorbidities, and PET), and we will consider multi-objective optimization techniques to select the base classifiers with the best balance between accuracy, diversity, and how easy to understand the motivating features behind the results are.

4. Conclusion

This paper proposed a pipeline for building an ensemble for AD progression detection based on multimodal time series data. The advantages of the proposed pipeline are twofold. *First*, it provides detailed steps to extract the most discriminating and medically critical features from static and time-series data. *Second*, it considers the trade-off between accuracy and diversity, which guides the selection of the optimum base classifiers. Based on the selected features from time series and static modalities, we created and compared many ensemble models based on either early data fusion and late decision fusion strategies. We conducted experiments using the ADNI dataset to tackle a 4-class classification problem based on 3-time series modalities and 1-static modality as inputs. The best-performing model is based on the stacking scheme with three XGB models at level-0 and the LR classifier as a meta classifier at level-1. The XGB models were fitted with three datasets, [CSs, NSB], [CSs, Static], and [NSB, Static], and the LR model is based on the class probabilities of the first-level classifiers. The experimental results showed that the proposed approach yields superior progression performance compared to all previous ensemble approaches in the literature. In summary, the proposed ensemble evaluates AD progression based on a cheap, easily collectible, simple set of features allowing the prediction of a patient's condition for 2.5 years in the future. In our future work, we will extend this work to consider other modalities, including the patient's comorbidities and drugs taken during the observation

period. In addition, we will investigate the interpretability of the proposed models to improve the medical expert's decision-making process.

Funding

Jose M. Alonso is Ramon y Cajal Researcher (RYC-2016–19802). This research is funded by the Spanish Ministry of Science, Innovation, and Universities (grants RTI2018-099646-B-I00 and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F2018/02, ED431C2018/29, ED431G/08, and ED431G2019/04). Some of the previous grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience Program (IITP-2021-2020-0-01821) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1011198).

6. Ethics approval

Data used in this study were obtained from the ADNI (https://adni.loni.usc.edu/). The Alzheimer's Disease Neuroimaging Initiative Data and Publications Committee (ADNI DPC) coordinates patient enrollment and ensures standard practice on the uses and distribution of the data as follows: The ADNI data were previously collected across 50 research sites. To participate in the study, each study subject gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites' Institutional Review Board (IRB). All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. A complete description of ADNI and up-to-date information is available at http://adni.loni.usc.edu/ and data access requests are to be sent to https://adni.loni.usc.edu/data-samples/access-data/. Detailed inclusion criteria for the diagnostic categories can be found at the ADNI website (https://adni.loni.usc.edu/methods). The ethics committees/institutional review board that approved the ADNI study are listed within Supplementary file (part 4).

Availability of data and material: The data are publicly available from the ADNI data repository (http://adni.loni.usc.edu). In addition, the specific patient RIDs used in our study and the full description of used features can be found in the Supplementary Files.

CRediT authorship contribution statement

Shaker El-Sappagh: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **Farman Ali:** Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **Tamer Abuhmed:** Conceptualization, Data curation, Formal analysis, Validation, Funding acquisition, Project administration, Writing - review & editing. **Jaiteg Singh:** Data curation, Validation, Writing – review & editing. **Jose M. Alonso:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration.

Data availability

The data that has been used is confidential.

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 data to this article can be found online at https://doi.org/10.1016/j.neucom.2022.09.009.

References

[1] K. Ritter, J. Schumacher, M. Weygandt, R. Buchert, C. Allefeld, J.D. Haynes, Multimodal prediction of conversion to Alzheimer's disease based onincomplete biomarkers, Alzheimer's Dement. Diagnosis, Assess. Dis. Monit. 1 (2) (2015) 206–215.
[2] H. Li, M. Habes, D.A. Wolk, Y. Fan, A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal MRI, Alzheimer's Dement. (2019) 1–12.
[3] R.C. Petersen et al., Mild cognitive impairment ten years later, Arch. Neurol. 66 (12) (2009) 1447–1455.
[4] 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.
[5] J. Ramírez et al., Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares, J. Neurosci. Methods 302 (2018) 47–57.
[6] D. Yao, V.D. Calhoun, Z. Fu, Y. Du, J. Sui, An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment, J. Neurosci. Methods 302 (2018) 75–81.
[7] E.E. Bron et al., Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, Neuroimage 111 (2015) 562–579.
[8] L. Sørensen et al., Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry, NeuroImage Clin. 13 (2017) 470–482.
[9] S.I. Dimitriadis, D. Liparas, Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) data, J. Neurosci. Methods 302 (2018) 14–23.
[10] G. Karas et al., Amnestic mild cognitive impairment: Structural MR imaging findings predictive of conversion to Alzheimer disease, Am. J. Neuroradiol. 29 (5) (2008) 944–949.
[11] A.K. Lebedeva et al., MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression, Front. Aging Neurosci. 9 (FEB) (2017) 1–11.
[12] 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.
[13] T. Grimmer et al., Visual versus fully automated analyses of 18F-FDG and amyloid PET for prediction of dementia due to Alzheimer disease in mild cognitive impairment, J. Nucl. Med. 57 (2) (2016) 204–207.
[14] O.V. Forlenza et al., Cerebrospinal fluid biomarkers in Alzheimer's disease: Diagnostic accuracy and prediction of dementia, Alzheimer's Dement. Diagnosis, Assess. Dis. Monit. 1 (4) (2015) 455–463.
[15] N. Mattsson, M. Ewers, K. Rich, E. Kaiser, E. Mulugeta, E. Rose, CSF Biomarkers and Incipient AlzheimerDisease in Patients With Mild Cognitive Impairment, JAMA 302 (4) (2009) 385–393.
[16] V. Youssofzadeh, B. McGuinness, L.P. Maguire, K. Wong-Lin, Multi-kernel learning with dartel improves combined MRI-PET classification of Alzheimer's disease in AIBL data: Group and individual analyses, Front. Hum. Neurosci. 11 (July) (2017) 1–12.
[17] M. Bucholc et al., A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual, Expert Syst. Appl. 130 (2019) 157–171.
[18] S. El-Sappagh et al., Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data, Futur. Gener. Comput. Syst. 115 (2021).
[19] R. Polikar et al., An ensemble based data fusion approach for early diagnosis of Alzheimer's disease, Inf. Fusion 9 (1) (2008) 83–95.
[20] E. Ruiz, J. Ramírez, J.M. Górriz, J. Casillas, Alzheimer's disease computer-aided diagnosis: Histogram-based analysis of regional mri volumes for feature selection and classification, J. Alzheimer's Dis. 65 (3) (2018) 819–842.

[21] T. Abuhmed, S. El-sappagh, J.M. Alonso, Robust hybrid deep learning models for Alzheimer's progression detection, Knowledge-Based Syst. 213 (2021) 106688.
[22] S. El-Sappagh, T. Abuhmed, S.M. 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.
[23] P.A. Donnelly-Kehoe, G.O. Pascariello, J.C. Gómez, Looking for Alzheimer's Disease morphometric signatures using machine learning techniques, J. Neurosci. Methods 302 (2018) 24–34.
[24] L. Nanni, A. Lumini, N. Zaffonato, Ensemble based on static classifier selection for automated diagnosis of Mild Cognitive Impairment, J. Neurosci. Methods 302 (2018) 42–46.
[25] C. Ledig, A. Schuh, R. Guerrero, R.A. Heckemann, D. Rueckert, Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database, Sci. Rep. 8 (1) (2018) 1–16.
[26] S. Iddi, D. Li, P.S. Aisen, M.S. Rafii, W.K. Thompson, M.C. Donohue, Predicting the course of Alzheimer's progression, Brain Informatics 6 (1) (2019).
[27] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Modeling disease progression via multi-task learning, Neuroimage 78 (2013) 233–248.
[28] W. Liu, B. Zhang, Z. Zhang, X.H. Zhou, Joint Modeling of Transitional Patterns of Alzheimer's Disease, PLoS One 8 (9) (2013).
[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.
[30] G. Gavidia-Bovadilla, S. Kanaan-Izquierdo, M. Mataroa-Serrat, A. Perera-Lluna, Early prediction of Alzheimer's disease using null longitudinal model-based classifiers, PLoS One 12 (1) (2017) 1–19.
[31] T. Wang, R.G. Qiu, M. Yu, Predictive Modeling of the Progression of Alzheimer ' s Disease with Recurrent Neural Networks, Sci. Rep. (2018) 1–12.
[32] S. El Sappagh, J.M. Alonso, S.M.R. Islam, A.M. Sultan, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Sci. Rep. (2021) 1–26.
[33] S. El-Sappagh, H. Saleh, F. Ali, E. Amer, T. Abuhmed, Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time, Neural Comput. Appl. (2022) 1–23.
[34] D. Jain, V. Singh, Feature selection and classification systems for chronic disease prediction: A review, Egypt. Informatics J. 19 (3) (2018) 179–189.
[35] M. Woźniak, M. Graña, E. Corchado, A survey of multiple classifier systems as hybrid systems, Inf. Fusion 16 (1) (2014) 3–17.
[36] W.L. Lee, An ensemble-based data fusion approach for characterizing ultrasonic liver tissue, Appl. Soft Comput. J. 13 (8) (2013) 3683–3692.
[37] K. Tumer, J. Ghosh, Analysis of decision boundaries in linearly combined neural classifiers, Pattern Recognit. 29 (2) (1996) 341–348.
[38] S. Farhan, M.A. Fahiem, H. Tauseef, An ensemble-of-classifiers based approach for early diagnosis of alzheimer's disease: Classification using structural features of brain images, Comput. Math. Methods Med. vol (2014) 2014.
[39] S. El-Sappagh, M. Elmogy, F. Ali, T. Abuhmed, S.M.R. Islam, K.-S. Kwak, A comprehensive medical decision–support framework based on a heterogeneous ensemble classifier for diabetes prediction, Electron. 8 (6) (2019).
[40] P.J. Moore, T.J. Lyons, J. Gallacher, Random forest prediction of Alzheimer's disease using pairwise selection from time series data, PLoS One 14 (2) (2019) 1–14.
[41] W. Yu, C. Zhao, Online fault diagnosis for industrial processes with bayesian network-based probabilistic, IEEE Trans. Autom. Sci. Eng. 16 (4) (2019) 1922–1932.
[42] A. Ebadi et al., Ensemble classification of Alzheimer's disease and mild cognitive impairment based on complex graph measures from diffusion tensor images, Front. Neurosci. 11 (FEB) (2017) 1–17.
[43] A.V. Lebedev et al., Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness, NeuroImage Clin. 6 (2014) 115–125.
[44] L. Sørensen, M. Nielsen, Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination, J. Neurosci. Methods 302 (2018) 66–74.
[45] S.H. Lee, D. Yu, A.H. Bachman, J. Lim, B.A. Ardekani, Application of fused lasso logistic regression to the study of corpus callosum thickness in early alzheimer's disease, J. Neurosci. Methods 221 (2014) 78–84.
[46] S. Cheriguene, N. Azizi, N. Dey, A.S. Ashour, A. Ziani, A new hybrid classifier selection model based on mRMR method and diversity measures, Int. J. Mach. Learn. Cybern. 10 (5) (2019) 1189–1204.
[47] R.S. Desikan et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (3) (2006) 968–980.
[48] G. McKhann D. Drachman M. Folstein R. Katzman D. Price E.M. Stadlan Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease Neurology 34 7 2012 939 939.
[49] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[50] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araújo, and J. Santos, "Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches," *ieee Comput. Intell. Mag.*, no. November, pp. 59–76, 2018.
[51] A.R.M. Forkan, I. Khalil, A clinical decision-making mechanism for context-aware and patient-specific remote monitoring systems using the correlations of multiple vital signs, Comput. Methods Programs Biomed. 139 (2017) 1–16.
[52] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79.
[53] Y. Perez-Riverol, M. Kuhn, J.A. Vizcaíno, M.P. Hitz, E. Audain, Accurate and fast feature selection workflow for high-dimensional omics data, PLoS One 12 (12) (2017) 1–14.
[54] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinformatics 1 (2015) 2015.
[55] S. Maldonado, R. Weber, F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, Inf. Sci. (Ny) 286 (2014) 228–246.
[56] R. Panthong, A. Srivihok, Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm, Procedia Comput. Sci. 72 (2015) 162–169.
[57] A.S. Rodin et al., Use of Wrapper Algorithms Coupled with a Random Forests Classifier for Variable Selection in Large-Scale Genomic Association Studies, J. Comput. Biol. 16 (12) (2010) 1705–1718.
[58] E. El-Houby, A survey on applying machine learning techniques for management of diseases, J. Appl. Biomed. 16 (3) (2018) 165–174.
[59] X.C. Yin, K. Huang, H.W. Hao, K. Iqbal, Z. Bin Wang, A novel classifier ensemble method with sparsity and diversity, Neurocomputing 134 (2014) 214–221.
[60] M. Aksela, J. Laaksonen, Using diversity of errors for selecting members of a committee classifier, Pattern Recognit. 39 (4) (2006) 608–623.
[61] E.K. Tang, P.N. Suganthan, X. Yao, An analysis of diversity measures, Mach. Learn. 65 (1) (2006) 247–271.
[62] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Is independence good for combining classifiers?, Proc - Int. Conf. Pattern Recognit. 15 (2) (2000) 168–171.
[63] D. Skalak, The sources of increased accuracy for two proposed boosting algorithms, Proc. Am. Assoc. Artif. Intell. (1996) 120–125.
[64] R.M.O. Cruz, R. Sabourin, G.D.C. Cavalcanti, Dynamic classifier selection: Recent advances and perspectives, Inf. Fusion 41 (2018) 195–216.
[65] S. Klöppel, A. Abdulkadir, C.R. Jack, N. Koutsouleris, J. Mourão-Miranda, P. Vemuri, Diagnostic neuroimaging across diseases, Neuroimage 61 (2) (2012) 457–463.
[66] Y. Klein-Koerkamp et al., Amygdalar Atrophy in Early Alzheimer's Disease, Current Alzheimer Research 11 (3) (2014) 239–252.
[67] O. Rémi Cuingnet, Emilie Gerardin, Jérôme Tessieras, Guillaume Auzias, Stéphane Lehéricy, Marie-Odile Habert, Marie Chupin, Habib Benali and Colliot, "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.
[68] S.F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J.C. Pruessner, D.L. Collins, Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning, Neuroimage 65 (2013) 511–521.
[69] C.R. Munteanu et al., Classification of mild cognitive impairment and Alzheimer's Disease with machine-learning techniques using 1 H Magnetic Resonance Spectroscopy data, Expert Syst. Appl. 42 (15–16) (2015) 6205–6214.
[70] D.P. Devanand et al., Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of Alzheimer disease, Neurology 68 (11) (2007) 828–836.
[71] H. Hampel, K. Bürger, S.J. Teipel, A.L.W. Bokde, H. Zetterberg, K. Blennow, Core candidate neurochemical and imaging biomarkers of Alzheimer's disease, Alzheimer's Dement. 4 (1) (2008) 38–48.
[72] C.A. Raji, O.L. Lopez, L.H. Kuller, O.T. Carmichael, J.T. Becker, Age, Alzheimer disease, and brain structure, Neurology 73 (22) (2009) 1899–1905.
[73] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, no. November 2018, pp. 1–12, 2019.
[74] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, A. Mueller, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[75] S. Whalen, G. Pandey, A comparative analysis of ensemble classifiers: Case studies in genomics, Proc. - IEEE Int. Conf. Data Mining, ICDM, 2013, pp. 807–816.
[76] H. Liu, L. Zhang, Advancing Ensemble Learning Performance through data transformation and classifiers fusion in granular computing context, Expert Syst. Appl. 131 (2019) 20–29.
[77] M. Saqlain, B. Jargalsaikhan, J.Y. Lee, A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing, IEEE Trans. Semicond. Manuf. 32 (2) (2019) 171–182.
[78] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, An ensemble learning approach for brain cancer detection exploiting radiomic features, Comput. Methods Programs Biomed. 185 (2020).
[79] C. Catal, M. Nangir, A sentiment classification model based on multiple classifiers, Appl. Soft Comput. J. 50 (2017) 135–141.
[80] M. Tanveer, A.H. Rashid, M.A. Ganaie, M. Reza, I. Razzak, K.-L. Hua, Classification of Alzheimer's disease using ensemble of deep neural networks trained through transfer learning, IEEE J. Biomed. Heal. Informatics (2021) 1.
[81] J. Liu, D. Zeng, R. Guo, M. Lu, F.X. Wu, J. Wang, MMHGE: detecting mild cognitive impairment based on multi-atlas multi-view hybrid graph convolutional networks and ensemble learning, Cluster Comput. 24 (1) (2021) 103–113.

[82] A. Giovannetti et al., Deep-MEG: spatiotemporal CNN features and multiband ensemble classification for predicting the early signs of Alzheimer's disease with magnetoencephalography, Neural Comput. Appl. 4 (2021).
[83] X. Bi, Y. Xie, H. Wu, L. Xu, Identification of differential brain regions in MCI progression via clustering-evolutionary weighted SVM ensemble algorithm, Front. Comput. Sci. 15 (6) (2021).
[84] M. N. K. P. and T. P., Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: A performance analysis Biomed. Signal Process. Control vol. 68, no. May 2021 102729.
[85] J. Ruiz, M. Mahmud, M. Modasshir, and M. Shamim Kaiser, "3D DenseNet Ensemble in 4-Way Classification of Alzheimer's Disease," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12241 LNAI, no. May 2021, pp. 85–96, 2020.
[86] N. An H. Ding J. Yang R. Au T.F.A. Ang Deep ensemble learning for Alzheimer's disease classification J. Biomed Inform. 105 May 2020 2019, p. 103411.
[87] 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 (May) (2020) 1–19.
[88] J.Y. Choi, B. Lee, Combining of Multiple Deep Networks via Ensemble Generalization Loss, Based on MRI Images, for Alzheimer's Disease Classification, IEEE Signal Process. Lett. 27 (2020) 206–210.
[89] A.H. Syed, T. Khan, A. Hassan, N.A. Alromema, M. Binsawad, A.O. Alsayed, An Ensemble-Learning Based Application to Predict the Earlier Stages of Alzheimer's Disease (AD), IEEE Access 8 (2020) 222126–222143.
[90] 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, Alzheimer's Dement. Diagnosis, Assess. Dis. Monit. 10 (2018) 737–749.
[91] M. Jin, W. Deng, Predication of different stages of Alzheimer's disease using neighborhood component analysis and ensemble decision tree, J. Neurosci. Methods 302 (2018) 35–41.
[92] P.J. Moore, J. Gallacher, T.J. Lyons, Random forest prediction of Alzheimer's disease using pairwise selection from time series data, arXiv:1808.03273 (2018) 1–6.
[93] R. Armañanzas, M. Iglesias, D.A. Morales, L. Alonso-Nanclares, Voxel-Based Diagnosis of Alzheimer's Disease Using Classifier Ensembles, IEEE J. Biomed. Heal. Informatics 21 (3) (2017) 778–784.
[94] L. Nanni, C. Salvatore, A. Cerasa, I. Castiglioni, Combining multiple approaches for the early diagnosis of Alzheimer's Disease, Pattern Recognit. Lett. 84 (2016) 259–266.
[95] T.R. Sivapriya, A.R.N.B. Kamal, P.R.J. Thangaiah, Ensemble Merit Merge Feature Selection for Enhanced Multinomial Classification in Alzheimer's Dementia, Comput. Math. Methods Med. (2015) 2015.
[96] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, Random forest-based similarity measures for multi-modal classi fi cation of Alzheimer ' s disease, Neuroimage 65 (2013) 167–175.
[97] I.J.A. Simpson, M.W. Woolrich, J.R. Andersson, A.R. Groves, J. Schnabel, Ensemble learning incorporating uncertain registration, IEEE Trans. Med. Imaging 32 (4) (2013) 748–756.
[98] M. Liu, D. Zhang, D. Shen, Ensemble sparse classification of Alzheimer's disease, Neuroimage 60 (2) (2012) 1106–1116.
[99] R. Chen et al., Prediction of conversion from mild cognitive impairment to Alzheimer disease based on Bayesian data mining with ensemble learning, Neuroradiol. J. 25 (1) (2012) 5–16.
[100] R. Polikar, C. Tilley, B. Hillis, C.M. Clark, Multimodal EEG, MRI and PET data fusion for Alzheimer's disease diagnosis, 2010 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC'10 (2010) 6058–6061.
[101] M. Silveira, J. Marques, Boosting Alzheimer disease diagnosis using PET images, Proc. - Int. Conf. Pattern Recognit. (2010) 2556–2559.
[102] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M.K. Chung, S.C. Johnson, Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset, Neuroimage 48 (1) (2009) 138–149.
[103] H. Gandhi, D. Green, J. Kounios, C.M. Clark, R. Polikar, Stacked generalization for early diagnosis of Alzheimer's disease, Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc. (2006) 5350–5353.

Shaker El-Sappagh received the bachelor's degree in computer science from Information Systems Department, Faculty of Computers and Information, Cairo University, Egypt, in 1997, and the master's degree from the same university in 2007. He received the Ph.D. degrees in computer science from Information Systems Department, Faculty of Computers and Information, Mansura University, Mansura, Egypt in 2015. In 2003, he joined the Department of Information Systems, Faculty of Computers and Information, Minia University, Egypt as a teaching assistant. Since June 2016, he has been with the Department of Information Systems, Faculty of computers and Information, Benha University where he is now an associated professor. He has publications in clinical decision support systems and semantic intelligence. His current research interests include machine learning, medical informatics, (fuzzy) ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, fuzzy expert systems, and cloud computing. He is a reviewer in many journals, and he is very interested in the diseases' diagnosis and treatment researches.

FARMAN ALI received the B.S. degree in computer science from the University of Peshawar, Pakistan, in 2011, the M.S. degree in computer science from Gyeongsang National University, South Korea, in 2015, and the Ph.D. degree in information and communication engineering from Inha University, South Korea, in 2018. He worked as a Postdoctoral Fellow at the UWB Wireless Communications Research Center, Inha University, from September 2018 to August 2019. He is currently an Assistant Professor with the Department of Software, Sejong University, South Korea. He has registered over four patents and published more than 50 research articles in peer-reviewed international journals and conferences. His current research interests include sentiment analysis/opinion mining, information extraction, information retrieval, feature fusion, artificial intelligence in text mining, ontology-based recommendation systems, healthcare monitoring systems, deep learning-based data mining, fuzzy ontology, fuzzy logic, and type-2 fuzzy logic. He has been awarded with the Outstanding Research Award (Excellence of Journal Publications-2017) and the President Choice of the Best Researcher Award during his Graduate Program at Inha University.

TAMER ABUHMED received the Ph.D. degree in information and telecommunication engineering from Inha University, in 2012. He is currently an assistant professor with the College of Computing and Informatics, Sungkyunkwan University, South Korea. His research interests include information security, expert and decision support systems, and machine learning based systems.

JAITEG SINGH is currently pursuing the Ph.D. degree in computer science and engineering with the Institutes of Higher Technical Education, with more than 15 years of experience in research, development, training, and academics. His areas of expertise are software engineering, business intelligence, data, opinion mining, cartography, curriculum design, and pedagogical innovation and management. His research interests include sustainable software engineering, education technology, offline navigation systems, and cloud computing.

JOSE M. ALONSO received the M.S. and Ph.D. degrees in telecommunication engineering from the Technical University of Madrid, Spain, in 2003 and 2007, respectively. He is currently a Post-Doctoral Researcher with the Research Centre in Information Technologies, University of Santiago de Compostela. He has published over 85 papers in international journals, book chapters, and peer-review conferences. His research interests include the development of free software tools, interpretable fuzzy modeling, natural language generation, knowledge extraction and representation, integration of expert and induced knowledge, data science, and big data. He is the Secretary of the European Society for Fuzzy Logic and Technology. He has been the Chair of the IEEE-CIS Task Force on Fuzzy Systems Software since 2018, and an Associate Editor of the IEEE Computational Intelligence Magazine.