Future Generation Computer Systems 115 (2021) 680–699

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data

Shaker El-Sappagh [a,b], Hager Saleh [c], Radhya Sahal [d], Tamer Abuhmed [e,*], S.M. Riazul Islam [f], Farman Ali [g,*], Eslam Amer [h,i]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain
[b] Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, Banha 13518, Egypt
[c] Faculty of Computers and Artificial Intelligence, South Valley University, Hurghada, Egypt
[d] Computer Science and Engineering College, Hodeidah University, Yemen
[e] Department of Computer Science and Engineering, College of Computing, Sungkyunkwan University, South Korea
[f] Department of Computer Science and Engineering, Sejong University, Seoul, South Korea
[g] Department of Software, Sejong University, Seoul, South Korea
[h] Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava, Czech Republic
[i] Faculty of Computer Science, Misr International University, Cairo, Egypt

ARTICLE INFO

Article history:
Received 26 February 2020
Received in revised form 8 September 2020
Accepted 5 October 2020
Available online 12 October 2020

Keywords:
Alzheimer disease
Machine learning
Multimodal data analysis
Disease progression detection

ABSTRACT

Alzheimer's disease (AD) is a severe neurodegenerative disease. The identification of patients at high risk of conversion from mild cognitive impairment to AD via earlier close monitoring, targeted investigations, and appropriate management is crucial. Recently, several machine learning (ML) algorithms have been used for AD progression detection. Most of these studies only utilized neuroimaging data from baseline visits. However, AD is a complex chronic disease, and usually, a medical expert will analyze the patient's whole history when making a progression diagnosis. Furthermore, neuroimaging data are always either limited or not available, especially in developing countries, due to their cost. In this paper, we compare the performance of five widely used ML algorithms, namely, the support vector machine, random forest, k-nearest neighbor, logistic regression, and decision tree to predict AD progression with a prediction horizon of 2.5 years. We use 1029 subjects from the Alzheimer's disease neuroimaging initiative (ADNI) database. In contrast to previous literature, our models are optimized using a collection of cost-effective time-series features including patient's comorbidities, cognitive scores, medication history, and demographics. Medication and comorbidity text data are semantically prepared. Drug terms are collected and cleaned before encoding using the therapeutic chemical classification (ATC) ontology, and then semantically aggregated to the appropriate level of granularity using ATC to ensure a less sparse dataset. Our experiments assert that the early fusion of comorbidity and medication features with other features reveals significant predictive power with all models. The random forest model achieves the most accurate performance compared to other models. This study is the first of its kind to investigate the role of such multimodal time-series data on AD prediction.

1. Introduction

Alzheimer's disease (AD) is the most severe form of dementia and often begins in people over the age of 65 [1]. According to the World Health Organization (WHO) [2], currently, there are 50 million people who have AD, and this number is expected to triple by 2050. Unfortunately, there is no cure for AD at this time, and current treatments can only reduce the future progression of the disease [3]. Early diagnosis of Alzheimer's disease (AD) is essential because AD treatment options tend to be most effective during the early stages of the disease [4]. Mild cognitive impairment (MCI) has been commonly viewed as a transitional stage between healthy aging and AD [5]. Studies have shown that 10%–15% of patients with MCI progress to AD per year [6]. A patient who converts from MCI to AD is called pMCI, whereas a patient who does not progress to AD is called sMCI [7]. One of the most difficult challenges is distinguishing between sMCI and pMCI. To date, many machine learning (ML) methods, such as support vector machine (SVM) have been applied to differentiate between sMCI and pMCI [8–10]. AD is a

* Corresponding authors.
E-mail addresses: tamer@skku.edu (T. Abuhmed), farmankanju@sejong.ac.kr (F. Ali).

https://doi.org/10.1016/j.future.2020.10.005
0167-739X/

chronic disease, where multiple modalities are always used to describe patients. These heterogeneous data collected over time can be referred to as time-series multimodalities [6]. The majority of AD diagnosis and progression studies are based on a single modality, usually neuroimaging data such as magnetic resonance imaging (MRI) [11]. Mostly, these studies often transform the AD diagnosis or prediction problem into a binary classification task (such as AD vs. MCI) to ease the training process. Bron et al. [12] organized the CADDementia[1] challenge to compare ML algorithms for AD diagnosis, 29 algorithms were evaluated based on single-visit MRI data only. The problem was formulated as a three-class classification problem (i.e. CN vs. MCI vs. AD). The best algorithm achieved an accuracy of 63.0% based on voxel-based morphometry features. Jiang et al. [13] built an interesting CNN-BN-DO-DA deep learning model for AD classification based on an eight-layer convolutional neural network with batch normalization and dropout techniques. This advanced model is based on the neuroimaging modality, and the authors achieved a high accuracy of 97.76% using a dataset of 7399 AD and 7399 normal subjects. Zhang et al. [14] proposed a novel machine learning system for automatic and fast AD diagnosis. This binary classifier is based on the volumetric MRI data of 196 subjects collected from two sources including the Open Access Series of Imaging Studies (OASIS) [15] and local hospitals (Affiliated Nanjing Brain Hospital of Nanjing Medical University, Children's Hospital of Nanjing Medical University, and Zhong-Da Hospital of Southeast University). The MRI data is processed by an accurate pipeline of skull stripping and spatial normalization, one axial slice, and stationary wavelet entropy. Based on the resulting texture features, a simple one-hidden layer neural network is used as the classifier, where the network parameters were trained using the particle swarm optimization. The resulting model is fast and achieved an accuracy of $92.73 \pm 1.03\%$. Although these models are highly accurate and promising, building an AD progression detection system based only on neuroimaging data is not highly recommended in the medical domain because real domain experts usually analyze complete patient profiles. Further, these imaging data are expensive to collect, which delays the AD diagnosis. In addition, according to standard AD clinical practice guidelines, neuroimaging is optional for AD diagnosis, and only required in specific situations like a history of carcinoma, bleeding disorders, and gait disorders; recent head trauma; age < 60 years; and rapid unexplained decline [16]. Recently, it has been proven that the fusion of multiple modalities improves the performance of the resulting models, where additional data such as position emission tomography (PET), neuropsychological battery, cognitive scores, symptoms, and demographics could enhance the model's confidence and reduce noise [17–20]. In addition, any resulting model becomes more acceptable in real medical environments. Zhang et al. [18] combined MRI, FDG-PET, and cerebrospinal fluid (CSF) modalities to distinguish AD, MCI, and normal controls patients. Xu, et al. [21] used the volumetric MRI, fluorodeoxyglucose PET (FDG-PET), and florbetapir PET modalities to classify AD vs. MCI in a binary classification task. Tong et al. [21] fused the volumetric MRI, voxel-based FDG-PET, CSF biomarker, and genetic modalities. Bouwman et al. [19] suggested incorporating the two modalities of MRI and CSF to distinguish CN patients from MCI. Gray et al. [22] used a random forest (RF) algorithm and four modalities (i.e., MRI, FDG-PET, CSF, and genetics) for the 3-class classification of AD vs. MCI vs. CN. All these studies were based only on the baseline data and did not study the role of time series data to enhance the classification process. In addition, they were based on advanced and expensive modalities such as MRI and PET. The technologies used to collect these data are unavailable in the majority of the medical clinics, which means that these classifiers are only applicable to limited patients. Furthermore, the results of using these modalities are not good. Donnelly-Kehoe et al. [23] concluded that the maximum accuracy achieved by MRI features did not reach that of using the mini-mental state examination (MMSE) alone.

Time-series data analysis is intuitive and crucial for the management of chronic diseases. However, in the AD domain, little work has used time-series algorithms for AD progression detection. In this context, Chincarini et al. [24] utilized a time-series MRI dataset from the Alzheimer's disease neuroimaging initiative (ADNI) to predict AD progression. These data have four scans (i.e. twice at baseline, one at 12-months, and one at 24 months). The study concentrated on analyzing the role of bilateral hippocampal volume to track AD progression. The problem was formulated as two binary classification tasks, and the study achieved an area under the ROC curve (AUC) of 0.93 for CN vs. AD and AUC of 0.88 for CN vs. MCI. Moradi et al. [7] predicted MCI-to-AD conversion in the period between one to three years based on novel MRI data using a semi-supervised learning technique. Moore et al. [25] used the random forest to study the relationship between pairs of data points at various time separations. Demographic, physical, and cognitive data were used to predict Alzheimer's disease. Huanget et al. [26] used a random forest regression algorithm to predict cognitive scores by utilizing the longitudinal scores of previous time points. To build accurate, stable, and medically intuitive models, multimodal time series data should be analyzed using suitable ML models. The usage of multimodal time series data for AD progression detection modeling is expected to improve model performance. In addition to MRI, PET, CSF, there are other crucial data sources, which are either have not been studied at all or have had few studies in the literature: (1) Cognitive score modalities like MMSE, CDRSB, FAQ, and ADAS 13 have only been studied at baseline, (2) drug modalities including brain disorders medications and other medications taken during the patient monitoring period, (3) comorbidity modalities which include the other diseases that the patient was suffering from during the monitoring period. These data have a great effect on a medical expert's decision to diagnose AD or predict its next stage [27]. For example, the ADNI collected drug modality determines the currently or previously taken medicines for the treatment of AD and other diseases [28–30]. These medicines have chemical substances that may be accumulating in the body in some form, so studying the effect of these drugs on the progression state of the disease is important. *However, to the best of our knowledge, these types of modalities have not been studied individually or in combination.*

In this study, we build a cost-effective and medically oriented AD progression detection system based on conventional ML techniques. The model is based on the information fusion of three-time series modalities of comorbidities, medications, and cognitive scores to predict four patient diagnosis classes: CN, AD, pMCI, and sMCI. In addition, basic demographics, including age, number of education years, and gender, are considered. Each modality is represented by four-time steps (i.e., baseline [bl], month 6 [M06], month 12 [M12], and month 18 [M18]), and the model predicts patient progression after 2.5 years (i.e., at month 48 [M48]). To select the optimum model, we optimize and test a set of five popular ML models, namely, SVM, RF, KNN, logistic regression (LR), and decision tree (DT), using the real world ADNI dataset. The preparation of medication and comorbidity datasets is a challenging task because the names of drugs seem to have been entered manually. Besides, there is a huge number of medications used by patients. Building a hot vector to encode these names created sparse datasets with many 0's. The resulting datasets are thus not suitable for ML algorithms. To semantically

[1] http://caddementia.grand-challenge.org.

Table 1
Patient statistics at baseline.

	CN (n = 249)	sMCI (n = 363)	pMCI (n = 106)	AD (n = 318)	Combined (n = 1036)
Gender (M/F)	144/105	210/153	44/62	142/176	483/553
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	00.28 ± 00.82	02.64 ± 03.31	07.63 ± 04.49	16.42 ± 06.59	06.81 ± 08.01
MMSE	28.91 ± 01.04	27.62 ± 01.95	25.46 ± 01.84	20.95 ± 03.95	25.66 ± 04.17
ADAS 13	08.13 ± 03.63	14.69 ± 06.71	22.69 ± 05.29	33.59 ± 09.39	19.73 ± 12.24

* Data are mean ± standard deviation.

manipulate these data, we utilized the semantics of the WHO's anatomical therapeutic chemical classification (ATC) ontology.[2] The drugs are grouped based on their chemical substances into a smaller number of classes, which significantly reduces the number of features used to encode drug data. The contributions of the paper can be summarized as follows.

- We propose a cost-effective, accurate, and medically intuitive AD progression detection model. The model is based on the early fusion of a set of new time series multimodalities to predict a 4-class classification task (i.e., CN, sMCI, pMCI, AD).
- We propose a novel methodology to encode the medication time-series data based on the standard ATC ontology.
- We implement, evaluate, and optimize a set of five models that are based on popular machine learning algorithms: SVM, RF, LR, DT, and KNN. These models are optimized to classify the 4-class problem (i.e., CN, sMCI, pMCI, AD), the 3-class task (CN, MCI, AD), and a set of binary classification tasks, including CN vs. AD, CN vs. MCI, sMCI vs. pMCI, etc.
- The models are trained and tested using real, time-series dataset of 1029 patients from the ADNI dataset.
- The results highlight the significant role of medication and comorbidity datasets to improve the performance of AD prediction models. The resulting models are more accurate compared to those in existing studies; besides, all models are less expensive because these models are based on well known efficient machine learning algorithms and are built using easy to collect and cheap historical data from patients.

The rest of this paper is structured as follows. Section 2 explains our methodology and the architecture of the proposed framework. Section 3 explains the experimental results, and Section 4 concludes the paper.

2. Materials and methods

In this section, we provide a little description of the ATC ontology, which has been used to encode the medication data. We also discuss the used cohort from the ADNI dataset. Furthermore, a detailed description of the proposed framework is discussed.

2.1. ATC ontology

ATC is a standard drug classification ontology established by the WHO [31] in 1976. This ontology is often used for the classification of active ingredients in drugs based on the organ on which they act and their therapeutic, pharmacological, and chemical properties. It has fifteen main anatomical/pharmacological groups. These groups are classified in hierarchies with five different levels (i.e., anatomical, therapeutic, pharmacological, chemical, and chemical substance). Each ATC main group is divided into 2 levels, which could be either pharmacological or therapeutic groups. The 3rd and 4th levels are chemical, pharmacological, or therapeutic subgroups, and the 5th level is the chemical substance. The 2nd, 3rd, and 4th levels are often used to identify pharmacological subgroups when that is considered more appropriate than therapeutic or chemical subgroups. This ontology is popular in the medical domain, where many research studies have used it to group drugs [27,32]. A detailed description of ATC can be found at www.whocc.no/atc/.

2.2. Cohort

We use a dataset obtained from the Alzheimer's disease neuroimaging initiative (ADNI) database with 1029 subjects who are categorized into four groups: (1) 246 subjects are CN at all time-points, (2) 362 subjects are sMCI at all time-points, (3) 105 subjects are pMCI, i.e., MCI at baseline + M06 + M12 + M18 visits and then convert to AD within 2.5 years from M18 (at M48), (4) 297 subjects are AD in all visits. Table 1 shows the statistics of the selected patients. The study is based on forecast-effective and medically critical time-series modalities (cognitive scores [CS], brain disorders medicines [AM], not brain disorders medicines [NAM], and comorbidities or disorders [D]), in addition to demographics or baseline data (B). The demographics include *age*, *gender*, *education years*. For each modality, we select the most popular features used in the literature that achieved the best results. The cognitive scores dataset has five features: *ADAS 13*, *CDG*, *FAQ, GDTOTAL*, and *NPISCORE*. Brain disorders medicines include seven features: *Aricept*, *Cognex*, *Exelon*, *Namenda*, *Razadyne*, and *Anti-Depressant*. These are the most common drugs taken by the studied patients. We added another feature to the AM, this feature is *None* to represent a patient who is not taking any drug. Not brain disorders medicines dataset includes 15 features which are encoded as *A*, *B*, *C*, *D*, *G*, *H*, *J*, *L*, *M*, *N*, *P*, *R*, *S*, and *V*. Another feature is added to NAM dataset which is *O* to represents other drugs taken. Note that these features represent level 1 of the ATC ontology. The disease dataset includes 17 features: *Psychiatric* (*MHPSYCH*), *Neurologic* (*MH2NEURL*), etc. The full description of the dataset features and codes can be found in Supplementary File 1.

2.3. Proposed framework

The architecture of the proposed model is shown in Fig. 1. It consists of the following components: data collection, data preprocessing, dataset splitting, dataset balancing, hyperparameter optimization, classifier training, model evaluation, and prediction. Our data has been collected from the ADNI dataset in the form of five time-series modalities. Next, we explain each component in detail.

2.3.1. Data preprocessing

In this step, we prepare the not brain disorders medicines, brain disorders medicines, and cognitive scores datasets.

[2] https://bioportal.bioontology.org/ontologies/ATC.

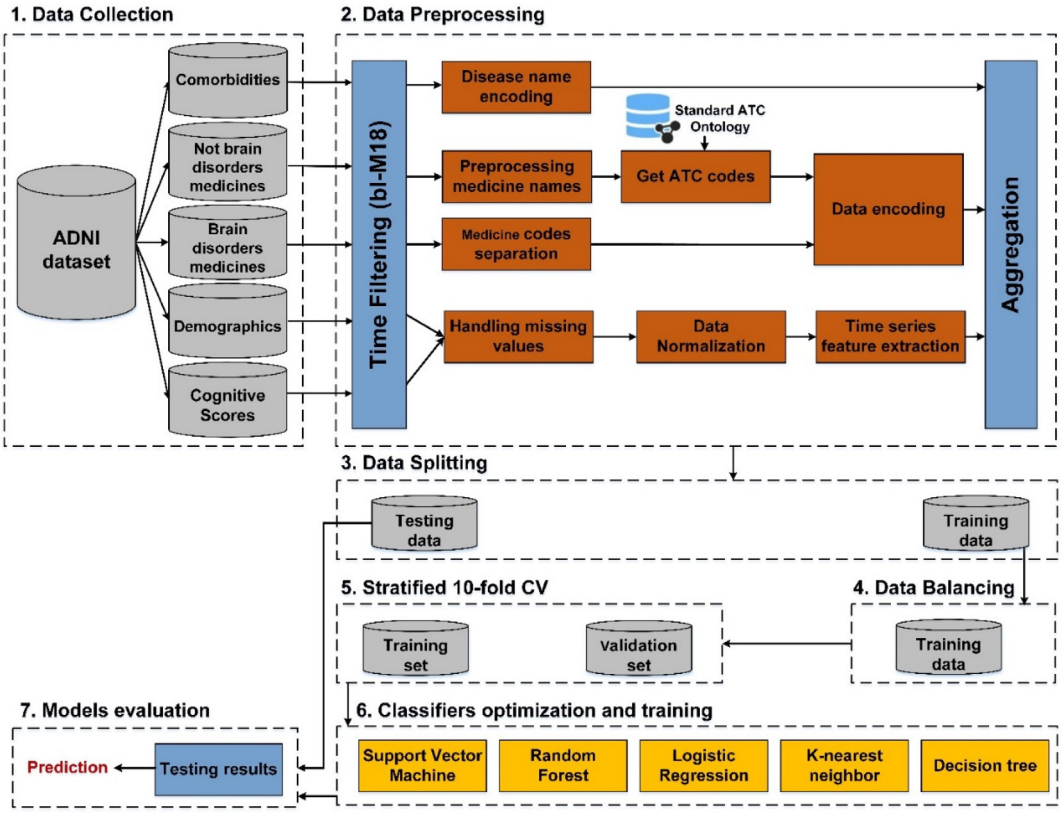

Fig. 1. The architecture of the proposed Alzheimer's progression prediction system.

2.3.1.1. Temporal data filtering. In this study, we examine the role of multimodal time-series data to predict the progression status of AD patients accurately. In each of the four time-series modalities (CS, D, AM, and NAM), a patient is represented as a collection of four rows, where each row is the description of the patient at that visit. In ADNI, these visit data are collected regularly every six months. ADNI collected data for more than ten years. However, the majority of these data are sparse and missing. In addition, these data are changing slowly, especially for stable patients (CN, sMCI, and AD). Based on statistical analysis, we found that the first four time-steps (i.e., bl, M06, M12, and M18) are the most complete and discriminative data. As a result, for each modality, a patient is represented using these four-time steps. These data summarize the patient's conditions within 18 months, and our model predicts the patient's progression status at M48, i.e., 2.5 years from M18. Fig. 2 shows the time series filtering process. The resulting data are used to extract statistical representative features, as discussed in Section 2.3.1.4.

2.3.1.2. Preparing the not brain disorders medicine data. Preparing the raw NAM dataset is a challenging task. The original dataset has over 5590 unique drug names. Creating a separate feature for each of these names results in a spare dataset, which is not suitable for learning. The dataset has so many abbreviations and drug names from different levels of abstraction. The main goals of this step are (1) clean drug names, (2) encode them by using standard medical oncology (ATC in our case), and (3) select the appropriate level of granularity for aggregating related drugs to ensure less sparse dataset.

• *Preparing drug name:* This step prepares the medical terminology of drug names to be suitable for searching in ATC ontology. All abbreviations have been changed to the full drug name, e.g. *ASA* and *Vit* have been replaced by *Aspirin* and *Vitamin*, respectively. Furthermore, some drugs did not use their original names, so these drug names are replaced with the drug class. For example, *Centrum*, which is considered as a type of vitamin, is replaced by *Vitamin*, and *Lexapro*, which is a type of escitalopram, is replaced by *Lexapro and Escitalopram*.

ATC encoding: The ATC standard ontology was used to (1) represent drug names in a standard format, (2) group related drugs under different categories to reduce the dimensionality of the resulting dataset. We propose the ATC medications encoding algorithm (AMD), which maps drug names into ATC codes (see Algorithm 1). The core function of the AMD procedure is calculating the similarity between the preprocessed medications' names and their corresponding ATC codes. The inputs of the AMD procedure are two datasets: the NAM dataset and ATC ontology. The ATC ontology is converted into a dataset of ⟨*ATC code*, *medicine name*⟩. All names are converted to lowercase. For the similarity measurement, we use a fuzzy string-matching similarity measure. This ratio of similarity between two strings is measured using the Levenshtein distance function.[3] The formal definition of the Levenshtein distance between two strings a and b is shown in Eq. (1).

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & if\ min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & otherwise. \end{cases} \quad (1)$$

where $1_{a_i \neq b_j}$ is the indicator function equal to 0 when $a_i \neq b_j$ and equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b. i and j are

[3] https://www.datacamp.com/community/tutorials/fuzzy-string-python.

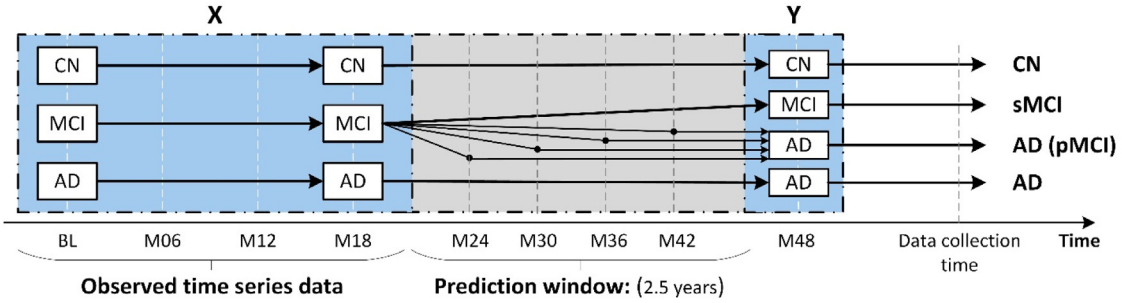

Fig. 2. Time series filtering process.

1-based indices. The fuzzy string-matching similarity (i.e., Levenshtein similarity ratio based on the Levenshtein distance) is implemented using the FuzzyWuzzyPython library, which gives the similarity index as a score from 0 to 100, where a score of 100 denotes the two strings are equal. After calculating the fuzzy similarity between drug names of the NAM and ATC ontology, we pick all the ATC codes that have similarity scores above 75 to be inserted in the resulting encoded dataset. The final step is to determine the level of granularity of the resulting codes. Using ontology semantics can dramatically decrease the number of features and keep the whole feature space knowledge. For example, according to the taxonomy of the ATC ontology, the drug names of Donepezil, Memantine, Ginkgo biloba, Tacrine, Rivastigmine, and Ipidacrine are types of Anticholinesterases, which is a Psychoanaleptic. As a result, we have replaced all of these subtypes by their parents in the ontology. The idea of semantically aggregating drug concepts according to formal ontology structure decreases the dimensionality and sparsity of the data. The resulting feature set of the AMD technique was at level 5 of the ATC and perfectly converted to 452 features. Aggregating drugs at the level 4 of ATC summarized the features set to 279 features, and aggregating them at the level 3 of the ATC summarized the features set to 160 features. Finally, aggregating the drugs at level 2 of the ATC summarized the features set to 76 features. This feature space was the most suitable because the number of features has been suitably reduced, and the data sparsity has been removed. The grouping of medicine names to levels 1 and 2 of ATC is achieved by removing the right digits from each code. Fig. 3 depicts level 1 and level 2 of the ATC. The final AMD database includes the ATC codes at level 1, which are *N*, *B*, *C*, *A*, *M*, *G*, *R*, *H*, *S*, *J*, *D*, *V*, *L*, and *P*. Other feature is *O* that represents the patient took another drug.

- *Data encoding:* In this step, a new dataset is created that has 15 columns, which represent level 1 of the ATC code. These columns are *N*, *B*, *C*, *A*, *M*, *G*, *R*, *H*, *S*, *J*, *D*, *V*, *L*, and *P*. We add other features to represent other medicines. Each column will be filled by 0 or 1. If the patient belongs to the group, the columns will be marked as 1; otherwise, 0.

```
Input: Not_Brain_Bisorders_Medicines [ ] = ['RID','Medicine_Name']
       Ontology_ATC_Dataset [ ] = ['Original_Medicine_Name','ATC Code']
Output: ATC_Decoded_Not_Brain_Bisorders_Medicines[ ] = ['RID','Medicine_Name','ATC_Code']
Step 1: Convert to Lowercases:
Not_Brain_Bisorders_Medicines.ToLowerCase('Medicine_ Name')
Ontology_ATC_Dataset.ToLowerCase('Medicine_Name')
Step 2: Similarity Calculation:
foreach i ∈ ATC_Decoded_Medicines_Dataset do
    foreach j ∈ Ontology_ATC_Dataset do
    Similarity_Ratio ⟵Fuzz.ratio(Not_Brain_Bisorders_Medicines[i]['Medicine_ Name'],
    Ontology_ATC_Dataset[j]['Original_Medicine_Name']
        if Similarity_Ratio > 75  then
          ATC_Decoded_Not_Brain_Bisorders_Medicines[i]['RID'] ⟵
          Not_Brain_Bisorders_Medicines[i]['RID']
          ATC_Decoded_Not_Brain_Bisorders_Medicines[i]['Medicine_Name'] ⟵
          Not_Brain_Bisorders_Medicines[i]['Medicine_ Name']
          ATC_Decoded_Not_Brain_Bisorders_Medicines[i]['ATC_Code'] ⟵
    Ontology_ATC_Dataset[j]['ATC_Code']
    end
end
```

2.3.1.3. Preparing the brain disorders medicines. The preprocessing of the brain disorders medicines dataset has the following steps.

- The brain disorders medicines dataset includes one column, which contains multiple ":" delimited values. Each value is defined in ADNI by a medication name. These values are separated into different columns using the one-hot encoding mechanism.
- The new dataset has eight columns: *Aricept*, *Cognex*, *Exelon*, *Namenda*, *Razadyne*, *Anti-depressant*, *Other*, and *None*. Each column will be filled by 1 if the patient takes the drug or 0 otherwise.

2.3.1.4. Preparing the cognitive scores and demographics. The preprocessing of the cognitive scores dataset has the following steps:

- *Handling the missing values*: For the demographics data, we first remove any feature with more than 30% missing. Next, we use the KNN algorithm to impute missing values, where missing values are replaced using the information from other subjects with a similar diagnosis. For CS, the scores that have more than 30% missing are removed. Patient cases with missing baseline scores were excluded. Time series values that are missing are handled using the following accurate procedure. If the diagnosis has not changed for a time step compared to its previous step, then we use forward filling with previous values. If the diagnosis has changed, we considered the value as missing and use the mean value according to each specific class (CN, sMCI, pMCI, and AD).
- *Data Normalization:* All numerical 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. This method converts data, so they have a 0 mean and unit standard deviation.
- *Time Series feature extraction:* From the CS, we collect one aggregated feature from the four historical time steps. For each patient, we collect the mean of each cognitive score. For patient x_i, the aggregated feature $\underline{s}_{t_1,t_2,\ldots,t_T} = (\underline{s}_1, \underline{s}_2, \ldots, \underline{s}_n)$, for vector $\underline{s}_{t_1,t_2,\ldots,t_T}$ with n dimensions, and $\underline{s}_j$ for $j = 1, 2, \ldots, n$ is the mean $(\frac{1}{N_{x_i}} \sum_{i=0}^{N_{x_i}} s_i)$, where in our case $N_{x_i} = 4$. The resulting value is expected to summarize the values of the four-time steps. After encoding the D, AM, and NAM modalities, using the hot vector representation, data summarization is achieved by collecting all the patient's drugs and comorbidities as a single row. This action is carried out by adding a value of 1 for the feature representing the drug *dr* or disease *di*, if the patient has taken the drug *dr* or is suffering from the disease *di* within the period of bl-M18. Please note that at each visit, the patient might be taking

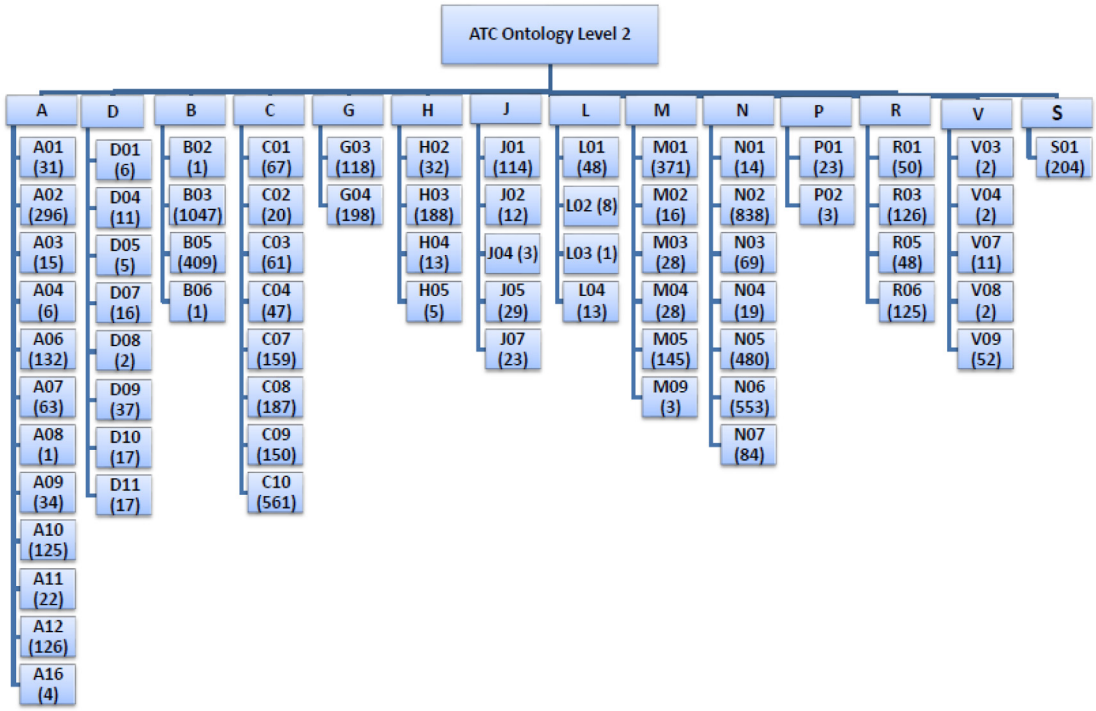

Fig. 3. ATC hierarchy using level 1 and level 2 groups.

multiple drugs or be suffering from multiple diseases. The hot vector representation of the ontology-based aggregated features facilitates the encoding process for these dynamic numbers representing the drugs and diseases.

2.3.2. Data fusion and splitting

In this step, the five modalities are fused, and the resulting dataset is split into training (90%) and testing sets (10%) using a stratified method. The training set is used to optimize and train the ML models, and the unseen test set is used to evaluate the resulting models.

2.3.3. Dataset balancing

Unbalanced datasets always result in biased results. To prevent this situation, the synthetic minority oversampling technique (SMOTE) oversampling [33] was utilized to handle the class imbalance. This technique was applied to the training set only. The testing set was balanced to mimic a real-world situation. The balanced training set was used to optimize and train the five machine learning methods based on a stratified 10-fold cross-validation (CV) mechanism.

2.3.4. Classifiers optimization and training process

The grid search method with a stratified 10-fold CV was used to find the optimal hyperparameters of all ML algorithms. The ML models tested in this study are SVM [34], DT [35], KNN [36], RF [37], and LR [38]. Each model is trained using the 90% training datasets created using the 10-fold CV techniques. Each experiment is repeated 10 times and the average is reported. The resulting models are evaluated using the unseen test set.

2.3.5. Evaluation metrics

To measure the CV and generalization performance of the proposed model, it must be based on a highly accurate and unbiased methodology. First, the data is randomly (and in a stratified way) divided into training (90%) and testing (10%) sets. This data splitting process is repeated 10 times and the average performance is recorded. In each splitting, the training set is balanced based on the popular SMOTE oversampling technique; then the balanced data are used to optimize the hyperparameters of the utilized ML techniques. Then, the stratified 10-fold CV technique is used to measure the CV performance. This stratified 10-fold CV process is repeated 10 times for every outer train/test split and the average performance is collected. Then, we trained the optimized model using the whole training set, next, we used the test set only to evaluate the performance of the final model. This strategy is very accurate and not biased because of the following:

(1) To correctly prepare the datasets, from the very beginning, we randomly (and in a stratified way) isolated a separate test set to be used for measuring the generalization performance of the ML models. This test set is not used in data normalization, and data balancing. The fitted normalization scaler is used to transform the testing sets. No data balancing is done before splitting to prevent repeating cases between training and testing sets. In addition, the selected features based on the training data were masked on the testing sets to filter these selected features.
(2) Model selection and hyperparameter optimization (based on the grid search technique) were performed based on the stratified 10-fold CV. Please note that the datasets used for this purpose have 926 cases, enough to measure stable performance in the evaluated models. In addition, this whole process was repeated 10 times and the average results are reported.
(3) Based on the discussion in (1), our data split prevents the mixing of model-selection and performance estimation, which supports the estimation of unbiased generalization performance in the models. As a result, the testing set measures the models' generalization performance. The data splitting process (i.e. training/test split) was repeated 10 times to measure stable testing performance.

Four standard metrics are used to evaluate the models, including accuracy, precision, recall, and F1-score, where TP is a true

positive, TN is a true negative, FP is a false positive, and FN is a false negative, see Eqs. (2)–(5).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

The non-parametric Friedman test and post-hoc Nemenyi test are used to measure the statistical significance of the differences in the performance of the 4-class and 3-class classifiers [39]. Please note that all binary classification classifiers achieved high performance, so to save space, we did not perform statistical analyses on all the binary classifiers. The Friedman test is a rank-based non-parametric equivalent of the repeated-measures ANOVA. It ranks the algorithms for each dataset separately, where the best one gets rank 1, the second-best rank 2, and so on. The statistics of the Friedman test are calculated, as shown in Eq. (6).

$$\chi_F^2 = \frac{12\text{D}}{\text{K}(\text{K}+1)} \left[\sum_{k=1}^{K} AvR_j^2 - \frac{K(K+1)^2}{4} \right] \tag{6}$$

where $AvR_j^2 = \frac{1}{D} \sum_{i=1}^{D} r_i^j$, D and K are the number of datasets and classifiers, respectively. r_i^j is the averaged rank of classifier j on dataset i. The null hypothesis is rejected only if the χ_F^2 is more significant than the critical value. This test only tells that there is a difference between classifiers, but cannot provide how a certain algorithm is statistically different. For that reason, the post-hoc Nemenyi test is used to perform paired comparisons, which defines a difference when the averaged ranks differ by at least a critical difference (CD). Critical values are calculated at the 0.05 significance level, and the F-Score metric is used to compare the models. The CD is calculated as shown in Eq. (7), where $q_{\alpha,\infty,\text{k}}$ is computed based on the t-test statistics.

$$\text{CD} = q_{\alpha,\infty,\text{k}} \sqrt{\frac{K(K+1)}{12D}} \tag{7}$$

3. Results and discussion

We examine the performance of the five ML models (i.e., RF, DT, LR, SVM, and KNN) based on different combinations of modalities. The main motivation for these early fusion strategies is to select the most informative and discriminative list of features and to highlight the role of different data types. Eight fused datasets were examined: (1) baseline and cognitive scores (B-CS), (2) baseline, cognitive scores, and brain disorders medicines (B-CS-AM), (3) baseline, cognitive scores, brain disorders medicines, and disease (B-CS-AM-D), (4) baseline, cognitive scores, brain disorders medicines, and not brain disorders medicines (B-CS-AM-NAM), (5) baseline, cognitive scores, brain disorders medicines, not brain disorders medicines, and disease (B-CS-AM-NAM-D), (6) baseline, cognitive scores, and disease (B-CS-D), (7) baseline, cognitive scores, not brain disorders medicines, and diseases (B-CS-NAM-D), and (8) baseline, cognitive score, and not brain disorders medicines (B-CS-NAM). The five ML classifiers were implemented using the scikit-learn 0.21.3 package in Python 3.7. Each of the eight datasets is split into a training set (90%) and a testing set (10%) using a stratified method. The training set is used to optimize the hyperparameters of the five models using the grid search technique. Stratified 10-fold cross-validation is used for hyperparameter tuning and model training. Each experiment was repeated six times, and the average result with standard deviation is reported. The testing set is used to evaluate the models on unseen data, which scores the generalization performance of the trained models. Three experiments were optimized to select the best list of features and the most accurate model. The following three sections discuss the collected results in detail.

Experiment 1 was used to optimize the ML models to solve the 4-class classification problem (i.e., CN, sMCI, pMCI, or AD). It has been asserted in the literature that this 4-class classification problem is difficult to optimize due to the similarity of the pMCI patients to AD patients. For instance, the pMCI patients are currently (i.e., from bl up to M18) MCI patients, but by M48, they will be AD patients. As a result, they confuse the ML classifiers because they are quite similar to sMCI from bl to M18, and quite similar to AD at M48. Our optimized 4-class models achieved optimum performance compared to previous literature. However, in Experiment 2, we tested the ML models on an easier 3-class classification task. In this experiment, we examined the optimized models on less complex classification problems that were either 3-class and 2-class. To convert the 4-class problem into a 3-class one, we considered pMCI cases as AD. Furthermore, the optimized models were then tested in Experiment 3 on a much easier binary classification task. We evaluated the models' performance in the binary classifications of AD vs. CN, AD vs. pMCI, and AD vs. sMCI.

3.1. Feature analysis

This study is based on a list of cost-effective features. These features are easy to obtain for all patients. As a result, the model is applicable in most medical domains, even in developing countries. The selected features, on the other hand, are medically intuitive and popular in the medical domain. Furthermore, they have discriminative and informative power to differentiate the four classes with high performance. For significance testing, the P-value is used, where 0.05 is the selected significance threshold. For CS displaying normal distributions, including ADAS and MoCA, we used a one-way analysis of variance (ANOVA) parametric test to check that the four independent classes are significantly different. For post-hoc testing after AVOVA, we use the conservative Scheffe test to check each pair of diagnostic groups. For the groups that had a non-normal distributions, including MMSE, FAQ, NPISCORE, GDTOTAL, and CDR, we used the non-parametric Kruskal-Wallis test for the three groups tests followed by Dunn's test for post-hoc multiple comparisons based on Bonferroni's corrections. The correlation between the categorical features, including drugs and comorbidities for all classes, was achieved using the Chi-Square test.

Regarding the cognitive scores data, only the most significant scores were selected. The sizes of different diagnostic groups (i.e., CN, sMCI, pMCI, and AD) are sufficient for all statistical tests, CS has no outliers. As shown in Fig. 4, these scores are statistically significantly different among the four classes ($P < .001$). As can be seen from the distributions in Fig. 4, there are significant differences among the means of the different classes for each score. As a result, the selected markers can be considered significant predictors of AD progression.

Regarding the list of brain disorders medicines, as shown in Fig. 5, *Aricept* is the most used drug among patients. In our dataset, 433 (42.18%) patients took it, with 236, 68, and 129 being from the AD, pMCI, and sMCI classes, respectively. *Namenda* is the next most important drug, where 265 (25.75%) patients took it, with 192, 37, and 36 being from the AD, pMCI, and sMCI classes, respectively. The third most important drug is *Anti-depressants*, where 247 (24.003%) patients took it, with 68, 52, 32, and 95 being from the AD, CN, pMCI, and sMCI classes, respectively. The

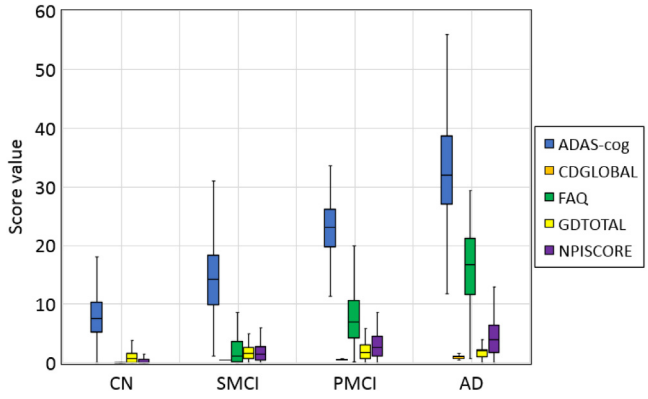

Fig. 4. Cognitive scores distributions.

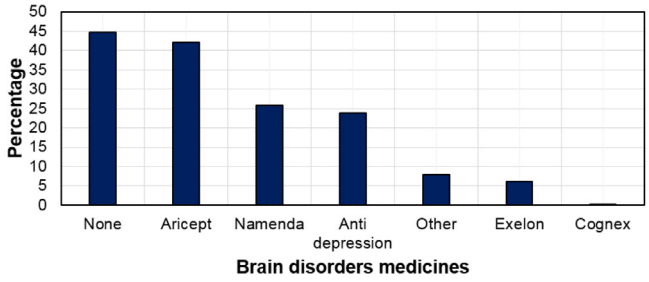

Fig. 5. Percentage of brain disorders medicines in the dataset.

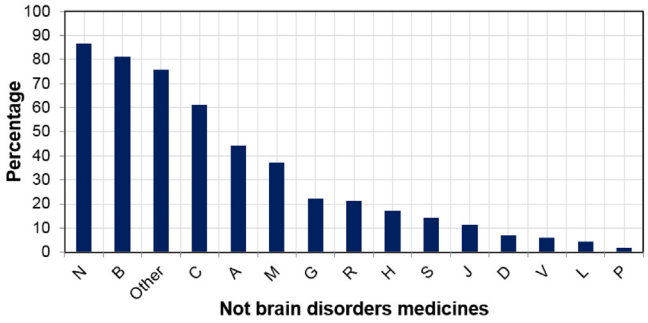

Fig. 6. Percentage of not brain disorders medicines in the dataset.

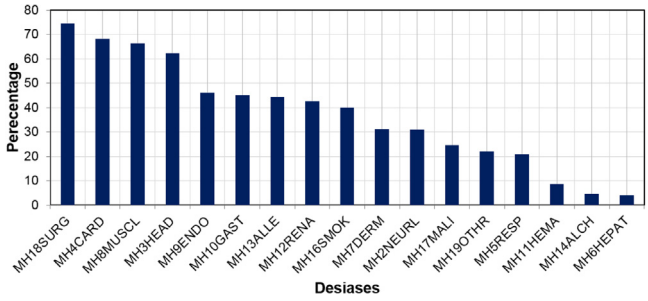

Fig. 7. Percentage of diseases in the dataset.

fourth most important drug is *Exelon*, where 64 (6.21%) patients took it, with 34 being from the AD class. *Razadyne* and *Cognex* are less used in AD society, where 31 (3.012%) and 1 (0.09%) patients took these drugs, respectively, and 16 and 1 patients were from the AD class, respectively. About 459 (44.60%) patients did not take any drugs (i.e., the *None* category), with 211 (45.97%) and 196 (36.82%) being from the CN and sMCI classes, respectively. Note that most of the normal patients do not take any drugs for brain disease. However, most AD patients have taken these kinds of drugs, while some pMCI and sMCI patients have taken them as well. The resulting features from the encoding of these drugs are statistically significant to discriminate different classes, except for *Cognex* ($P > .5$) and *Razadyne* ($P > .17$), which as a result are excluded from the dataset.

Regarding the not brain medicine drugs, as shown in Fig. 6, the *N* is the most common group in the ATC; see Supplementary File 1. In our dataset, 892 (86.77%) patients took it, with 278 of them being from the AD class. *B* is the second most important group, where 836 (81.32%) patients took it, with 252 of them being from the AD class. The third most important group is *C*, where 631 (61.28%) patients took it, with 200 of them being from the AD class. The fourth most important group is *A*, where 454 (41.16%) patients took it, with 130 of them being from the AD class. The fifth most important group is *M*, where 353 (37.35%) patients took it, with 99 of them being from the AD class. The sixth most important group is *G*, where 229 (22.27%) patients took it, with 63 of them being from the AD class. The seventh most important group is *R*, where 220 (21.40%) patients took it, with 62 of them being from the AD class. The eighth most important group is *H*, where (177) 17.21% patient took it, with 39 of them being from the AD class. The ninth most important group is *S*, where 147 (14.29) patients took it, with 36 of them being from the AD class. *J*, *D*, *V*, *L*, *P* are less popular groups, where 118 (11.47%), 73 (7.10%), 60 (5.83%), 46 (4.47%) and 19 (1.84%) patients took these drugs, respectively, with 24, 12, 14, 8 and 7 of them being from the AD class, respectively. Finally, around 779 (75.77%) patients took other drugs, and these patients are grouped under the *Other* category. These features are statistically significant, according to the Chi-Square Test.

Regarding the patient's medical history, as shown in Fig. 7, 766 (74.61) of the patients did have *major surgical procedures* (MH18SURG), with 389 of them being from the AD class. About 701 (74.5%) patients had *cardiovascular disorders* (MH4CARD), with 376 of them being from the AD class. *Musculoskeletal disorders* (MH8MUSCL) is the next most important disease, where 683 (66.43%) patients experienced that, with 370 of them being from the AD class. The fourth most important group had medical issues with the head, eyes, ears, nose, or throat (MH3HEAD), where 640 (62.25%) patients experienced that, with 349 of them being from the AD class. The fifth most important group had *endocrine system diseases* (MH9ENDO), and 474 (46.10%) patients experienced that, with 264 of them being from the AD class. The sixth most important group had medical issues with *gastrointestinal disorders* (MH10GAST), where 465 (45.23%) patients experienced that, with 238 of them being from the AD class. *Allergies* (MH13ALLE) and *renal-genitourinary* (MH12RENA) include 457 (44.45%) and 437 (42.50%) patients, respectively, with 228 and 223 being from the AD class, respectively. *Smoking* group (MH16SMOK) has 413 (40.17%) patients, with 225 being from the AD class. The *malignancy*(MH17MALI) and *respiratory disorders* (MH5RESP) have 252 (24.51%), and 214 (20.817%) patients, respectively, with 134, 118, and 109 of them being from the AD class. H*ematopoietic lymphatic disorders* (MH11HEMA), *alcohol abuse* (MH14ALCH), and *hepatic disorders* (MH6HEPAT) were less common, with 90 (8.75%), 47 (4.57%) and 43 (4.18%) patients, respectively. These features were also found to be statistically significant following the Chi-Square Test.

3.2. The 4-class experiment

3.2.1. Cross-validation results

This section discusses the 10-fold CV results of the five models over the eight training datasets (i.e., B-CS, B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, B-CS-D, B-CS-NAM-D, and

Table 2
Performance of ML models in the 4-class task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	87.85	87.85	87.85	87.85	87.02 ± 2.83	87.23 ± 2.80	86.75 ± 2.96	86.90 ± 2.87
	B-CS-AM	90.13	90.13	90.13	90.13	90.11 ± 2.54	90.48 ± 2.42	90.13 ± 2.72	90.23 ± 2.48
	B-CS-AM-D	**90.51**	**90.69**	**90.51**	**90.41**	90.58 ± 2.68	90.95 ± 2.37	90.48 ± 2.54	90.54 ± 2.51
	B-CS-AM-NAM	90.16	90.09	90.16	90.09	**90.89 ± 2.73**	**91.26 ± 2.55**	**91.07 ± 2.74**	**90.94 ± 2.62**
	B-CS-AM-NAM-D	89.70	89.82	89.70	89.58	90.78 ± 2.72	91.05 ± 2.78	90.68 ± 2.52	90.62 ± 2.54
	B-CS-D	89.47	89.45	89.46	89.38	89.43 ± 2.21	89.99 ± 1.95	89.50 ± 2.09	89.58 ± 2.11
	B-CS-NAM-D	89.35	89.32	89.35	89.28	90.33 ± 2.15	90.57 ± 2.00	90.35 ± 2.21	90.17 ± 2.18
	B-CS-NAM	88.65	88.75	88.65	88.59	88.76 ± 2.32	89.28 ± 2.32	88.93 ± 2.34	88.86 ± 2.47
DT	B-CS	72.92	72.92	72.92	72.92	**78.94 ± 3.54**	**80.76 ± 3.81**	**79.17 ± 4.57**	**79.98 ± 3.90**
	B-CS-AM	**77.32**	**77.44**	**77.31**	**77.05**	78.43 ± 4.46	79.49 ± 4.52	79.28 ± 4.03	78.02 ± 5.49
	B-CS-AM-D	68.28	68.37	68.28	67.77	72.94 ± 5.48	74.52 ± 5.46	73.63 ± 5.88	72.35 ± 5.23
	B-CS-AM-NAM	75.70	75.79	75.69	75.31	77.66 ± 4.30	78.04 ± 4.14	77.29 ± 4.33	78.38 ± 4.65
	B-CS-AM-NAM-D	72.45	72.45	72.45	72.17	73.47 ± 4.61	73.60 ± 5.07	73.81 ± 5.58	74.18 ± 4.86
	B-CS-D	75.58	75.91	75.57	75.48	76.57 ± 4.15	76.60 ± 4.28	76.26 ± 4.29	76.06 ± 4.84
	B-CS-NAM-D	71.64	71.87	71.64	71.56	72.31 ± 6.15	72.51 ± 6.33	72.85 ± 5.73	72.34 ± 5.02
	B-CS-NAM	76.39	76.72	76.39	76.31	73.55 ± 5.89	77.06 ± 4.79	75.12 ± 4.65	75.64 ± 4.27
LR	B-CS	79.40	79.40	79.40	79.40	74.33 ± 3.44	73.96 ± 3.72	74.34 ± 3.44	73.57 ± 3.58
	B-CS-AM	84.83	84.81	84.83	84.77	76.47 ± 3.21	76.26 ± 3.38	76.45 ± 3.22	76.03 ± 3.33
	B-CS-AM-D	**85.53**	**85.63**	**85.53**	**85.32**	78.65 ± 2.82	78.73 ± 2.92	78.66 ± 2.82	78.45 ± 2.92
	B-CS-AM-NAM	84.49	84.29	84.49	84.29	77.59 ± 2.60	77.48 ± 2.83	77.59 ± 2.69	77.28 ± 2.77
	B-CS-AM-NAM-D	84.37	84.57	84.37	84.10	**79.01 ± 3.44**	**78.98 ± 3.57**	**79.03 ± 3.45**	**78.72 ± 3.55**
	B-CS-D	81.48	81.63	81.48	81.27	75.85 ± 2.66	75.71 ± 2.80	75.85 ± 2.66	75.28 ± 2.67
	B-CS-NAM-D	80.67	80.88	80.67	80.32	76.57 ± 2.27	76.38 ± 2.55	76.57 ± 2.27	76.12 ± 2.49
	B-CS-NAM	79.39	79.50	79.39	79.23	75.37 ± 2.17	75.11 ± 2.31	75.36 ± 2.17	74.80 ± 2.28
SVM	B-CS	79.63	79.63	79.63	79.63	83.01 ± 2.97	83.21 ± 3.05	83.01 ± 2.97	82.52 ± 3.13
	B-CS-AM	81.59	82.24	81.59	81.47	84.94 ± 2.79	85.04 ± 2.82	84.94 ± 2.79	84.54 ± 2.99
	B-CS-AM-D	82.63	82.63	82.63	82.46	86.47 ± 2.59	86.68 ± 2.56	86.47 ± 2.59	86.17 ± 2.69
	B-CS-AM-NAM	82.06	82.10	82.06	81.83	**87.32 ± 2.23**	**87.61 ± 2.25**	**87.32 ± 2.23**	**87.04 ± 2.29**
	B-CS-AM-NAM-D	**83.68**	**83.51**	**83.68**	**83.52**	86.13 ± 3.03	86.37 ± 3.03	86.13 ± 3.03	85.85 ± 3.07
	B-CS-D	81.71	81.64	81.71	81.47	86.02 ± 2.41	85.95 ± 2.45	86.03 ± 2.41	85.69 ± 2.51
	B-CS-NAM-D	82.40	82.13	82.40	82.16	86.47 ± 2.38	86.47 ± 2.43	86.47 ± 2.38	86.21 ± 2.48
	B-CS-NAM	80.78	80.76	80.78	80.48	84.74 ± 2.08	84.58 ± 2.26	84.74 ± 2.08	84.36 ± 2.19
KNN	B-CS	71.52	71.52	71.52	71.53	72.12 ± 3.81	74.34 ± 4.23	72.12 ± 3.81	70.74 ± 3.94
	B-CS-AM	74.30	79.80	74.30	72.93	74.63 ± 3.25	77.63 ± 3.65	74.63 ± 3.25	73.03 ± 3.48
	B-CS-AM-D	71.18	74.42	71.18	69.97	72.92 ± 3.52	75.61 ± 3.89	72.92 ± 3.52	71.33 ± 3.80
	B-CS-AM-NAM	71.41	78.04	71.41	69.41	73.50 ± 3.21	76.98 ± 3.76	73.5 ± 3.21	71.63 ± 3.68
	B-CS-AM-NAM-D	71.52	74.70	71.52	70.43	71.46 ± 3.72	73.49 ± 4.26	71.46 ± 3.72	69.88 ± 3.99
	B-CS-D	**75.69**	**79.45**	**75.69**	**74.94**	**76.03 ± 2.27**	**77.93 ± 2.52**	**76.03 ± 2.27**	**74.55 ± 2.44**
	B-CS-NAM-D	75.46	79.51	75.46	74.61	75.88 ± 2.71	77.76 ± 2.96	75.88 ± 2.71	74.33 ± 2.90
	B-CS-NAM	71.87	77.06	71.87	70.18	73.55 ± 3.06	75.69 ± 3.23	73.55 ± 3.06	71.82 ± 3.44

B-CS-NAM), as shown in Table 2, where X-Y represents modality X is fused with modality Y. RF is the best model among all fused datasets. Accuracies for B-CS, B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, B-CS-D, B-CS-NAM-D, and B-CS-NAM were 87.02%, 90.11%, 90.58%, 90.89%, 90.78%, 89.43%, 90.33%, and 88.76%, respectively.

Furthermore, the B-CS-AM-NAM dataset has the highest performance with respect to other fused datasets. We attribute this behavior to a fusion of the Alzheimer's medicine dataset and the Not Alzheimer's medicine dataset, which play an important role in improving the performance of all ML models. We experimentally demonstrate the performance improvement of RF using different fused datasets compared to the B-CS dataset. As the results show, RF accuracy improvement for B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, and B-CS-NAM-D datasets is +4%; for B-CS-AM and B-CS-D datasets it is +3%, and for B-CS-NAM dataset it is +2%. RF has recoded improvements in precision and F1-score of 4% for B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-D, and B-CS-NAM-D datasets, and +3% and +2% for B-CS-D and B-CS-NAM datasets, respectively. Furthermore, RF recorded a higher recall of +5% by using the B-CS-AM-NAM dataset. Note that RF achieves the highest results by using the B-CS-AM-NAM dataset.

For other ML models, their performance varied based on the utilized datasets, see Table 2. Regarding the B-CS dataset, the worst model is KNN (accuracy of 72.12%, precision of 74.34%, recall of 72.12%, and F1-score of 70.74%). Regarding the B-CS-AM dataset, the worst model is KNN (accuracy of 74.63%, precision of 77.63%, recall of 74.63%, and F1-score of 73.03%). Regarding the B-CS-AM-D dataset, the worst model is KNN (accuracy of 72.92%, precision of 75.61%, recall of 72.92%, and F1-score of 71.33%). Regarding the B-CS-AM-NAM dataset, the worst model is KNN (accuracy of 71.46%, precision of 73.49%, and F1-score of 69.88%). Regarding the B-CS-AM-NAM-D dataset, the worst model is DT (accuracy of 73.46%, precision of 73.60%, and F1-score of 74.18%). Regarding the B-CS-D dataset, the worst model is LR (accuracy of 75.84%, precision of 75.71%, recall of 75.84%, and F1-score of 75.28%). For the B-CS-AM-NAM dataset, the worst model is LR (accuracy of 75.48%, precision of 75.71%, and F1-score of 75.28%). Regarding the B-CS-NAM-D dataset, the worst model is DT (accuracy of 72.31%, precision of 72.51%, and F1-score of 72.34%). Finally, DT is the worst model for the B-CS-NAM dataset (accuracy of 72.31%, precision of 72.51%, and F1-score of 72.34%). Although KNN recorded the lowest performances using B-CS-AM, B-CS-AM-D, and B-CS-AM-NAM datasets, the trained model showed improvements compared to using the B-CS dataset (accuracy of +3%, +1%, and +2% respectively). LR recorded lower performances using B-CS-AM and B-CS-AM-D datasets comparing to RF, but the model showed improvements compared to using the B-CS dataset (accuracy of +6.84% and +7.72% respectively).

According to Table 2, LR has the highest performance improvements using the B-CS-AM-NAM-D dataset compared to the B-CS dataset. However, we notice that DT recorded the lowest performance using B-CS-AM-NAM-D and B-CS-NAM datasets, and its

performance is lower compared to the B-CS dataset. Interestingly, for the B-CS-D dataset, KNN has lower performance than LR, but it has a higher performance improvement compared to the B-CS dataset (accuracy of +5%). However, KNN failed to improve its performances using the B-CS-AM-NAM-D dataset (accuracy of −1%, precision of −1%, recall of −1%, and F1-score of −1%). DT achieves the worst performance on the selected modalities using all fused datasets compared to the baseline dataset, which means that DT is unable to predict AD progression using four classes with the training datasets. Compared to RF, SVM reports lower performances when using the B-CS-AM-NAM dataset, but it has a higher performance improvement compared to using the B-CS dataset (accuracy of +5%).

Fig. 8 shows the performance of the best performing models based on the selected datasets. These results are the average of the results collected from ten repeated train/test splits wherein each splitting of the training set is performed with stratified 10-fold CV. It is clear that RF achieves the best performance (accuracy of 90.89%) by using the B-CS-AM-NAM fused dataset followed by SVM (accuracy of 87.32%) using the B-CS-AM-NAM dataset, LR (accuracy of 79.01%) using the B-CS-AM-NAM-D dataset, DT (accuracy of 78.94%) using the B-CS, and then KNN (accuracy of 76.03%) using the B-CS-D dataset. Generally, the fused datasets were more complex than single modalities, and therefore the classification tasks became more challenging. It can be noticed that the complex models like RF ensemble and SVM are able to utilize the fused datasets to enhance their performance, but lazy learners like KNN achieved worse results. This indicates that these models benefit from added variances in the data and are able to avoid the effect of added noise. From Table 2, we noticed that the RF-based model is the most stable because it has the lowest standard deviations compared to other models. To conclude, it can be seen from Table 2 that the medication data has an important role in improving the performance of ML models when fused with other modalities.

3.2.2. Testing results

This section discusses the generalization performance of the five models using the eight unseen test datasets, see Table 2. The results are consistent with cross-validation performances. RF is the best model among all other ML models based on the B-CS-AM-D dataset (accuracy of 90.51%). The fusion of more modalities to the B-CS dataset improves the performance of RF. Its accuracy is improved by 3% using the B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, B-CS-D, or B-CS-NAM-D datasets, and 1% using the B-CS-NAM dataset. A similar improvement is achieved for precision, recall, and F1-score, where RF records improvements of 3% for B-CS-AM and B-CS-AM-D; 2% for B-CS-AM-NAM, B-CS-AM-D, B-CS-NAM-D, and B-CS-D; and 1% for B-CS-NAM. DT achieves high performance using B-CS-AM, B-CS-AM-NAM, and B-CS-NAM datasets, compared to B-CS. However, it has lower performance using B-CS-AM-D, B-CS-AM-NAM-D, and B-CS-NAM-D, compared to the baseline dataset. LR achieves higher performance compared to B-CS for all fused datasets expect B-CS-NAM. SVM does a good job using the B-CS-AM-NAM-D dataset. The greatest recall improvements were made by KNN using B-CS-AM, B-CS-AM-NAM, B-CS-D, B-CS-NAM-D, and B-CS-NAM (10%, 8%, 10%, 10%, and 7%, respectively). KNN improved precision by 10%, 4%, 8%, 4%, and 7% using the respective B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, and B-CS-NAM datasets. However, it has worse F1-scores of −2%, −3%, −2%, and −2% for B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, and B-CS-NAM, respectively, compared to the B-CS dataset.

Fig. 9 shows the performance of the best models from each category. To collect the test results, after each split, all models were tested using the same test set. The collected testing results

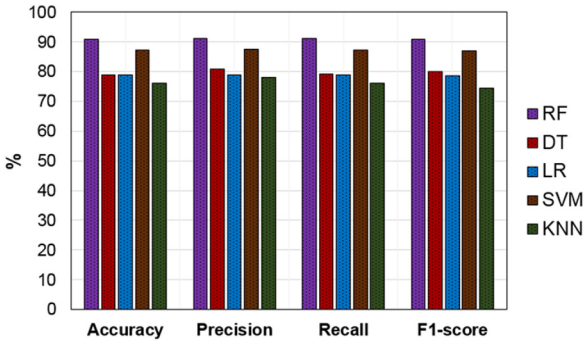

Fig. 8. The best CV performance for 4-class models.

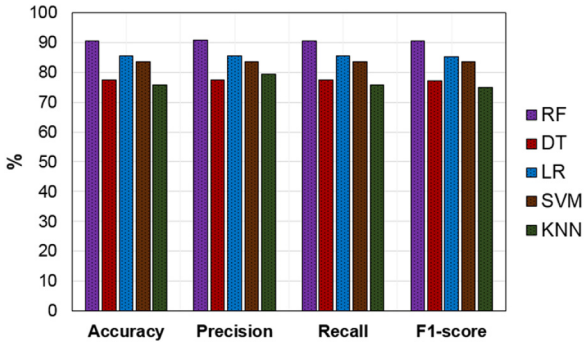

Fig. 9. The best generalization results for 4-class models.

from the ten splits were averaged and are shown in Fig. 9. As a result, every model has been tested 10 times, and every testing split has been used with every single model. Again, RF achieved the best testing results by utilizing the B-CS-AM-D dataset (accuracy of 90.51%). LR achieved better testing performance than SVM based on the B-CS-AM-D dataset (85.53%). It can be noticed that all models utilized the medication and comorbidities datasets to improve their results. This result shows how much these types of data may add value to the classification performance. SVM utilized B-CS-AM-NAM-D to achieve an accuracy of 83.68%. Please note that this level of performance has not been achieved in previous literature based on just the MRI data. DT achieved an accuracy of 77.32% using B-CS-AM, and finally, KNN achieved an accuracy of 75.69% based on the B-CS-D dataset. All techniques have better results based on the fused data. More complex models such as SVM and RF utilized more modalities and benefited from the added variances of the new features.

To summarize the performance of the compared models, we explore the average testing results of each model over all datasets. On average, RF achieved an average testing performance of Accuracy = 89.48%, Precision = 89.51%, Recall = 89.48%, and F1-Score = 89.41%. By these results, RF achieved the highest average results over all datasets with a 95%-confidence interval (CI) of [81.40, 94.10]. LR achieved the second-best results over all datasets (Accuracy = 82.52%, Precision = 82.59%, Recall = 82.52%, and F1-Score = 82.34%). SVM achieved the third best average testing results (Accuracy = 81.81%, Precision = 81.83%, Recall = 81.81%, and F1-Score = 81.63%). DT and KNN achieved the lowest average results, where DT has Accuracy = 73.79%, Precision = 73.93%, Recall = 73.78%, and F1-Score = 73.57%, and KNN has Accuracy = 72.87%, Precision = 76.81%, Recall = 72.87%, and F1-Score = 71.75%. Fig. 10 illustrates a comparison between the average performance of all models using the eight test sets. Fig. 10(a) confirms that the RF ensemble classifier outperformed all other regular ML models, and simple models like DT or lazy

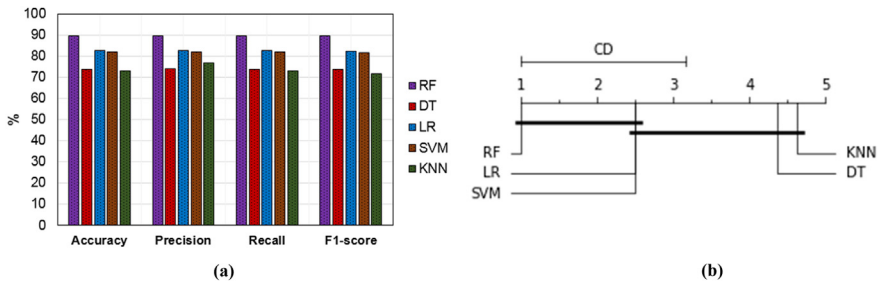

Fig. 10. (a) Average performance over the eight test sets for 4-class models, (b) Critical difference diagram.

models like KNN were unable to benefit from the fused feature space. The CD diagram, shown in Fig. 10(b), asserts that RF achieved the lowest average ranking, statistically outperforming SVM, DT, LR, and KNN. LR and SVM achieved lower average rank when compared to DT and KNN. The reason for KNN obtaining the highest average rank is due to the fact that it always obtains a lower rank for most datasets.

3.3. The 3-class experiment

In this experiment, we test the role of multimodal data fusion on the performance of ML models to solve less complex classification tasks (i.e., 3-class problem). We examine the previously optimized ML models using different fusion schemes.

3.3.1. Cross-validation results

This section discusses the 10-fold CV results of the five models over the eight training datasets using three classes, as shown in Table 3. Generally, all models have higher performance for the 3-class problem compared to the 4-class problem. RF is the best model among all tested models. Its accuracies for B-CS, B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, B-CS-D, B-CS-NAM-D, and B-CS-NAM are 90.86%, 91.58%, 90.81%, 92.04%, 91.35%, 89.82%, 90.08%, and 90.01%, respectively. The B-CS-AM-NAM dataset has the highest performance. RF accuracy improvement for B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, and B-CS-NAM-D datasets is 4%, for B-CS-AM and B-CS-D datasets is 3% and for B-CS-NAM dataset is 2%.

Similar to the precision and F1-score, RF has recoded improvements with respect to B-CS, 4% for B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-D, and B-CS-NAM-D datasets, 3% and 2% for B-CS-D and B-CS-NAM datasets respectively. However, for RF recall improvement, it achieves 5% using the B-CS-AM-NAM dataset. We attribute this behavior to a fusion of Alzheimer's medicine dataset and Not Alzheimer's medicine dataset, which has an essential role in improving the performance of all ML models. KNN has the worst performances. Using the B-CS dataset, the model achieved an accuracy of 72.12%, a precision of 74.34%, a recall of 72.12%, and an F1-score of 70.74%. It also has the worst results using the B-CS-AM dataset (accuracy of 78.27%, precision of 78.64%, recall of 78.27%, and F1-score of 77.56%). However, KNN has gained significant improvements using other fused datasets. For the DT model, early fusion slightly improved the model's performance. For example, the performance was improved by using the B-CS-AM-NAM-D and B-CS-D datasets by accuracies of 3% and 2%, respectively. However, DT performance was not improved by using the B-CS-NAM-D, and B-CS-NAM datasets. SVM, on the other hand, has gained performance improvements using all the fused datasets. LR has superior improvements for all fused datasets. It gained the highest performance improvements using the B-CS-AM dataset (i.e. precision and F1-score of 17%).

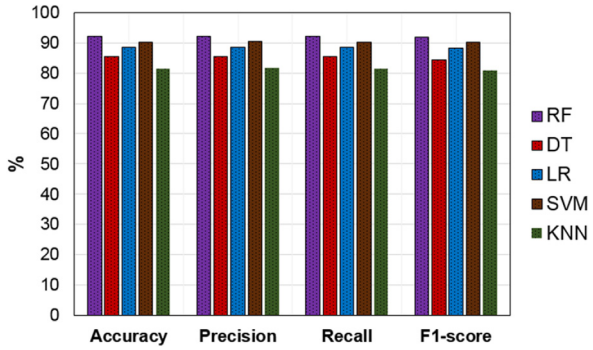

Fig. 11. The best CV performance for 3-class models.

3.3.2. Testing results

Table 3 shows the performance of the ML models using the unseen testing datasets. RF is the best performing model, where both B-CS-AM-D and B-CS-AM-NAM datasets have the highest accuracy (91.21%) and recall (91.21%). However, B-CS-AM-NAM has higher precision and F1-score. On the other hand, DT has the lowest performances using all unseen datasets, where B-CS-NAM-D achieved the lowest performances (accuracy of 63.69%, precision of 63.75%, recall of 63.69%, and F1-score of 63.37%). KNN saw the best improvement based on fused datasets in comparison to the B-CS dataset. It improved in accuracy by 12%, 16%, 17%, and 13% for B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, and B-CS-AM-NAM-D, respectively, but RF achieved 3%, 4%, 4%, and 4%, respectively for the same datasets. Furthermore, KNN achieved higher F1-scores by 18%, 17%, and 19% for B-CS-AM, B-CS-AM-NAM, and B-CS-AM-NAM-D datasets, respectively. The same behavior was followed by LR, where fused datasets improved its results compared to the baseline data. LR models improved their results by 10% for accuracy, precision, recall, and F1-score for both B-CS-AM and B-CS-D datasets. In addition, using B-CS-AM-NAM-D datasets with SVM improved performance by 10%. Even for DT, which has the worst performances, it achieved slightly improved accuracy using the B-CS-AM, B-CS-AM-D and B-CS-AM-NAM-D datasets by 3%, 2%, and 1%, respectively.

Fig. 11 shows a comparison among the best models for every algorithm based on the CV results, and Fig. 12 is based on the testing results. The CV and testing results were collected using the same methodology used in the 4-class problem. All models have better performance compared to the 4-class task. This is intuitive because the current task is much easier than the previous experiment. As illustrated in the figure, RF beats all other techniques using the fused dataset of B-CS-AM-NAM (accuracy of 92.05%). Furthermore, the model is more stable because its results have lower variance compared to other models. SVM utilized B-CS-AM to get an accuracy of 90.25%.

LR and DT used the same dataset to achieve accuracies of 88.41% and 85.57%, respectively. KNN achieved the worst performance (81.31%) based on the B-CS-AM-NAM dataset. RF behaves,

Table 3
Performance of ML models in the 3-class task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	87.26	87.17	87.26	87.17	90.86 ± 2.49	90.84 ± 2.15	90.88 ± 2.29	90.85 ± 2.29
	B-CS-AM	90.05	90.07	90.05	90.01	91.58 ± 2.22	91.78 ± 2.04	91.48 ± 2.42	91.73 ± 2.03
	B-CS-AM-D	91.21	91.19	91.21	91.10	91.82 ± 2.06	91.83 ± 2.27	91.62 ± 2.21	91.74 ± 2.26
	B-CS-AM-NAM	**91.21**	**91.20**	**91.21**	**91.19**	**92.05 ± 2.31**	**92.29 ± 2.37**	**92.04 ± 2.44**	**92.01 ± 2.34**
	B-CS-AM-NAM-D	90.95	90.94	90.95	90.92	91.35 ± 2.69	91.75 ± 2.45	91.44 ± 2.63	91.49 ± 2.61
	B-CS-D	89.40	89.39	89.40	89.37	89.83 ± 2.57	90.08 ± 2.55	89.67 ± 2.67	90.02 ± 2.75
	B-CS-NAM-D	89.40	89.38	89.40	89.38	90.08 ± 2.61	90.31 ± 2.60	90.03 ± 2.77	90.11 ± 2.58
	B-CS-NAM	87.38	87.44	87.38	87.32	90.01 ± 2.22	90.36 ± 2.28	90.16 ± 2.29	90.39 ± 2.11
DT	B-CS	75.69	76.17	75.69	75.56	78.90 ± 3.97	79.97 ± 3.96	78.62 ± 4.31	79.05 ± 4.30
	B-CS-AM	**77.90**	**77.83**	**77.9**	**77.68**	**85.57 ± 4.17**	**85.54 ± 4.48**	**85.36 ± 3.87**	**84.37 ± 3.86**
	B-CS-AM-D	77.51	77.42	77.51	77.26	83.33 ± 3.85	82.76 ± 4.46	83.31 ± 3.91	83.54 ± 4.09
	B-CS-AM-NAM	76.39	76.72	76.39	76.31	83.11 ± 4.39	84.69 ± 4.37	82.13 ± 4.02	84.60 ± 3.38
	B-CS-AM-NAM-D	72.22	72.02	72.22	71.95	81.32 ± 6.36	82.05 ± 5.77	81.09 ± 5.77	81.70 ± 5.36
	B-CS-D	69.76	69.94	69.76	69.04	80.44 ± 5.14	81.52 ± 3.83	80.10 ± 5.86	80.63 ± 6.03
	B-CS-NAM-D	63.69	63.75	63.69	63.37	78.58 ± 5.48	78.22 ± 5.37	77.11 ± 6.56	79.39 ± 5.12
	B-CS-NAM	67.36	67.39	67.36	67.10	77.76 ± 4.26	78.30 ± 4.48	76.95 ± 3.93	78.44 ± 4.43
LR	B-CS	79.40	79.45	79.40	79.34	74.34 ± 3.44	73.95 ± 3.72	74.36 ± 3.43	73.58 ± 3.55
	B-CS-AM	**88.37**	**88.51**	**88.37**	**88.41**	**88.41 ± 2.79**	**88.63 ± 2.81**	**88.41 ± 2.79**	**88.32 ± 2.88**
	B-CS-AM-D	86.95	87.04	86.95	86.98	87.97 ± 2.82	88.17 ± 2.86	87.97 ± 2.82	87.86 ± 2.91
	B-CS-AM-NAM	87.72	87.76	87.72	87.72	88.01 ± 2.44	88.17 ± 2.47	88.01 ± 2.44	87.89 ± 2.45
	B-CS-AM-NAM-D	87.72	87.76	87.72	87.72	87.82 ± 2.50	87.93 ± 2.50	87.82 ± 2.50	87.74 ± 2.53
	B-CS-D	87.98	88.001	87.98	87.96	85.78 ± 2.83	85.93 ± 2.84	85.78 ± 2.83	85.70 ± 2.83
	B-CS-NAM-D	85.91	86.06	85.91	85.95	85.50 ± 2.55	85.61 ± 2.68	85.50 ± 2.55	85.39 ± 2.60
	B-CS-NAM	79.39	79.50	79.39	79.23	75.38 ± 2.18	75.12 ± 2.32	75.38 ± 2.17	74.81 ± 2.29
SVM	B-CS	79.63	82.46	79.63	79.71	83.01 ± 2.97	83.21 ± 3.05	83.01 ± 2.97	82.52 ± 3.13
	B-CS-AM	86.04	85.96	86.04	85.97	**90.25 ± 2.31**	**90.61 ± 2.32**	**90.25 ± 2.31**	**90.27 ± 2.31**
	B-CS-AM-D	87.21	87.11	87.21	87.08	89.80 ± 2.46	90.01 ± 2.45	89.79 ± 2.46	89.78 ± 2.47
	B-CS-AM-NAM	85.91	86.03	85.91	85.82	89.68 ± 2.31	89.89 ± 2.25	89.68 ± 2.31	89.67 ± 2.29
	B-CS-AM-NAM-D	**87.69**	**87.69**	**87.69**	**87.69**	88.29 ± 2.55	88.54 ± 2.50	88.29 ± 2.55	88.28 ± 2.55
	B-CS-D	84.36	84.66	84.36	84.46	87.76 ± 2.98	88.15 ± 2.84	87.76 ± 2.98	87.81 ± 2.93
	B-CS-NAM-D	85.53	85.78	85.53	85.61	86.37 ± 2.91	86.82 ± 2.77	86.37 ± 2.91	86.42 ± 2.86
	B-CS-NAM	80.78	80.76	80.78	80.48	84.74 ± 2.08	84.58 ± 2.26	84.74 ± 2.08	84.36 ± 2.19
KNN	B-CS	71.52	74.59	71.52	70.35	72.12 ± 3.81	74.34 ± 4.23	72.12 ± 3.81	70.74 ± 3.94
	B-CS-AM	81.13	81.96	81.13	85.97	78.27 ± 3.18	78.64 ± 3.33	78.27 ± 3.18	77.56 ± 3.28
	B-CS-AM-D	85.27	85.61	85.27	80.83	80.17 ± 2.85	80.56 ± 2.87	80.17 ± 2.85	79.71 ± 2.86
	B-CS-AM-NAM	**86.56**	**86.83**	**86.56**	**85.18**	**81.31 ± 2.97**	**81.75 ± 2.96**	**81.32 ± 2.97**	**80.87 ± 3.17**
	B-CS-AM-NAM-D	82.30	82.90	82.30	86.52	78.07 ± 3.24	78.47 ± 3.48	78.07 ± 3.24	77.28 ± 3.50
	B-CS-D	82.30	82.77	82.30	82.03	77.60 ± 3.33	77.56 ± 3.63	77.60 ± 3.33	76.91 ± 3.51
	B-CS-NAM-D	82.04	82.48	82.04	82.08	77.77 ± 3.18	77.75 ± 3.37	77.77 ± 3.18	77.09 ± 3.33
	B-CS-NAM	71.88	77.06	71.88	81.87	73.55 ± 3.06	75.69 ± 3.23	73.55 ± 3.06	71.82 ± 3.44

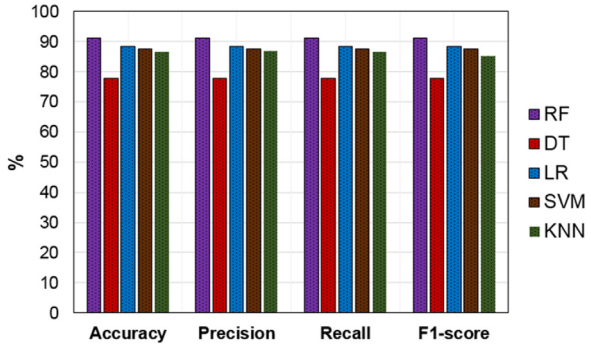

Fig. 12. The best generalization results for 3-class models.

in the same way, using the testing datasets. It achieved the highest accuracy of 91.21% based on B-CS-AM-NAM. However, LR achieved better accuracy than SVM (accuracy of 88.37%) based on the B-CS-AM dataset. SVM has an accuracy of 87.69% based on the B-CS-AM-NAM-D dataset. DT has the lowest results (77.9%) based on the B-CS-AM dataset. It can be noticed that a combination of AM and NAM modalities with the baseline data has a great effect on model performance. These results highlighted the main role of a patient's cost-effective history data on improving AD progression detection models.

For each 3-class classifier, we explore the average testing results over all datasets. On average, RF achieved the highest average testing performance of Accuracy = 89.61%, Precision = 89.60%, Recall = 89.61%, and F1-Score = 89.56%, with a 95%-confidence interval (CI) of [81.50, 94.20]. LR achieved the second-best results over all datasets (Accuracy = 85.43%, Precision = 85.51%, Recall = 85.43%, and F1-Score = 85.41%). SVM achieved the third highest performance in its average testing results (Accuracy = 84.64%, Precision = 85.06%, Recall = 84.64%, and F1-Score = 84.60%). KNN had average performance of Accuracy = 80.38%, Precision = 81.78%, Recall = 80.38%, and F1-Score = 81.85%. DT achieved the lowest average results of Accuracy = 72.57%, Precision = 72.66%, Recall = 72.57%, and F1-Score = 72.28%. Fig. 13 illustrates a comparison between the average performance of all models using the eight test sets. As shown in Fig. 13(a), the RF ensemble classifier outperformed all other regular ML models. The CD diagram of Fig. 13(b) shows that RF statistically outperforms SVM, DT, LR, and KNN as it achieves the lowest average ranking. LR and SVM achieved lower average performance when compared to the DT and KNN models, while DT achieved the worst results.

3.4. The 2-class experiment

In this section, we test the optimized ML models using the three binary classification tasks of AD vs.CN, AD vs. sMCI, AD vs. pMCI, and CN vs. MCI.

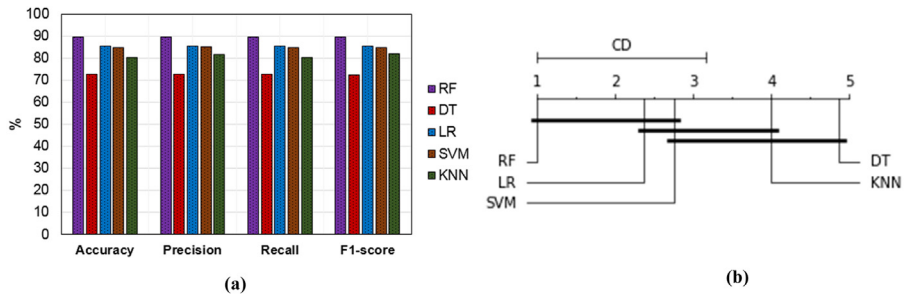

Fig. 13. (a) Average performance over the eight test sets for 3-class models, (b) Critical difference diagram.

3.4.1. Cross-validation results: AD vs. CN

In this subsection, we discuss the results of the 10-fold CV for the AD vs. CN task using the five models over eight training datasets, see Table 4. It should be noted that all ML models have high performance over all trained datasets with results ranging between 98.28% and 99.77% for accuracy, 99.63% and 99.77% for precision, 96.945% and 99.74 for recall, and 97.025% and 99.74% for F1-score. RF has the highest performance using all training sets. It achieves the highest result by using the B-CS-AM dataset (accuracy of 99.71%, precision of 99.77%, recall of 99.65% and F1-score of 99.68%). Compared to the B-CS dataset, RF has the highest improvement in performance (accuracy by+0.24% using B-CS-D and B-CS-NAM-D, precision by+0.14% using B-CS-AM, recall by+0.12% using B-CS-NAM-D, and F1-score by+0.15% using B-CS-NAM-D). For the DT model, the best results were achieved using the B-CS-NAM dataset (accuracy of 96.77%, precision of 96.99%, recall of 96.945%, and F1-score of 97.03%). DT achieved small improvements (accuracy by+0.02% and F1-score by +0.52%) using the B-CS-AM-NAM dataset. Similarly, LR had a small improvement using all fused datasets compared to the baseline dataset. Its highest improvements were achieved by using B-CS-AM-NAM and B-CS-AM-NAM-D datasets (0.18%, 0.19%, 0.18%, and 0.18% for accuracy, precision, recall, and F1-score, respectively). Some fused sets positively affect the performance of the KNN, but others did not. Its highest improvement was achieved using the B-CS-AM-D dataset (i.e. 0.11%, 0.11%, 0.16%, and 0.11% for accuracy, precision, recall, and F1-score respectively). The SVM model showed superior performance improvements. The B-CS-AM-NAM-D dataset achieved the highest results with SVM (i.e.+0.90%). The CN class has different characteristics compared to the AD class. As a result, it is an easy task for all ML models when looking at this class. As a result, there is a minor difference in results among models. Besides, there is no big difference in performance by adding more features to the training sets.

3.4.2. Testing results: AD vs. CN

Table 4 shows the performance of ML models using the unseen data for the AD vs. CN task. RF achieved again the best results of 100%. SVM has similar performance except for B-CS-AM-D and B-CS-AM-NAM-D datasets. KNN, LR, and DT have accuracies of 98.44%, 99.74, and 97.40, respectively. It should be noted that some models like RF and SVM achieved 100% performance using most datasets. Both models achieved this performance using the B-CS dataset. That is because the AD vs. CN classification is an easy task. On the other hand, some models like KNN, DT, and LR did not achieve this level of performance, adding more features did not improve these models' performance.

3.4.3. Cross-validation results: AD vs. sMCI

This section discusses the results of the 10-fold CV for the AD vs. sMCI classification task, see Table 5. The sMCI class is more similar to the CN class. Despite this, all ML models were still able to discriminate patients with high accuracy. All results are slightly lower than for the previous AD vs. CN task because the sMCI patients are more similar to AD patients than CN. The performance of the ML models ranged between 88.55% and 94.91% for accuracy, 88.78% and 95.1% for precision, 87.92% and 94.91% for recall, and 88.67% and 94.96% for F1-score. The RF model has the highest performance, while DT has the lowest performance.

Compared to the B-CS dataset results, RF is the only model that achieved performance improvements for all fused datasets. The highest improvement was obtained by using the B-CS-AM-NAM dataset (i.e. +1.61%, +1.77%, +1.66%, and +1.70% for accuracy, precision, recall, and F1-score respectively). LR only obtained performance improvement using the B-CS-AM-NAM-D dataset (i.e., 0.06%, 0.03%, 0.06%, and 0.06% for accuracy, precision, recall, and F1-score, respectively). Similarly, SVM and KNN only obtained performance improvements using the B-CS-AM dataset (i.e. 0.54% and 0.03% for accuracy, 0.49% and 0.03 for precision, 0.54%, and 0.03% for recall, and 0.54% and 0.03% for F1-score, respectively). DT did not show any performance improvement using any of the fused datasets.

3.4.4. Testing results: AD vs. sMCI

Table 5 shows the testing performance of the ML models for the AD vs. sMCI task. The results assert that the AD vs. sMCI task is more challenging than the AD vs. CN task, which is medically intuitive. For RF, the highest accuracy and recall is 92.36% using the B-CS-AM dataset with a 2% improvement compared to the B-CS dataset. Its highest precision is 92.19% by using the B-CS-AM-NAM dataset with a 2% improvement, and its highest F1-score is 92.35% using the B-CS-AM dataset. By using B-CS-AM-D and B-CS-NAM-D datasets, LR achieved these results for accuracy (89.12%), precision (89.12%), recall (89.12%), and F1-score (89.12%). Compared to the baseline performance, the results improved by 2.8% for recall and 2.9% for accuracy, precision, and F1-score. SVM achieved the best improvement (8%) by using the B-CS-AM-NAM-D dataset. KNN recorded an improvement of 2% by using the B-CS-AM-D, B-CS-AM-NAM, and B-CS-NAM-D datasets. Finally, DT recorded a performance improvement of 3.1% and 0.3% for accuracy, precision, recall, and F1-score by using B-CS-D and B-CS-AM-NAM, respectively. It can be noticed that adding more modalities generally improves the model's performance and stability. We noticed that adding more data has less effect when the task is easy for a classifier. For example, on average, the models improved performance by 8% in the case of the AD vs. sMCI task compared to the easier AD vs. CN task.

3.4.5. Cross-validation results: AD vs. pMCI

In this experiment, we evaluate the optimized ML models to differentiate AD from pMCI. Although this is a binary classification task, it is more difficult than the previous two experiments. The pMCI patients are quite similar to AD patients. They have converted from MCI to AD within 2.5 years from M18. As a result, the models performance seen in Table 6 reflects the level of challenge

Table 4
Performance for the AD vs. CN task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	100.0	100.0	100.0	100.0	99.50 ± 0.50	99.63 ± 0.37	99.62 ± 0.38	99.59 ± 0.41
	B-CS-AM	100.0	100.0	100.0	100.0	99.71 ± 0.30	99.77 ± 0.23	99.65 ± 0.35	99.68 ± 0.32
	B-CS-AM-D	100.0	100.0	100.0	100.0	99.65 ± 0.35	99.66 ± 0.34	99.65 ± 0.35	99.68 ± 0.32
	B-CS-AM-NAM	100.0	100.0	100.0	100.0	99.62 ± 0.38	99.61 ± 0.39	99.71 ± 0.29	99.56 ± 0.41
	B-CS-AM-NAM-D	100.0	100.0	100.0	100.0	99.48 ± 0.52	99.69 ± 0.31	99.53 ± 0.47	99.59 ± 0.41
	B-CS-D	100.0	100.0	100.0	100.0	99.74 ± 0.26	99.73 ± 0.27	99.70 ± 0.30	99.73 ± 0.27
	B-CS-NAM-D	**100.0**	**100.0**	**100.0**	**100.0**	**99.74 ± 0.26**	**99.73 ± 0.27**	**99.74 ± 0.26**	**99.74 ± 0.26**
	B-CS-NAM	100.0	100.0	100.0	100.0	99.70 ± 0.30	99.66 ± 0.34	99.61 ± 0.39	99.70 ± 0.30
DT	B-CS	96.61	96.65	96.61	96.61	98.63 ± 1.46	98.74 ± 1.53	98.86 ± 1.22	98.54 ± 1.64
	B-CS-AM	96.87	96.98	96.87	96.87	98.50 ± 1.50	97.99 ± 2.01	98.19 ± 1.08	98.74 ± 1.55
	B-CS-AM-D	96.61	96.65	96.61	96.61	98.31 ± 1.93	98.55 ± 1.48	98.34 ± 1.66	98.10 ± 1.90
	B-CS-AM-NAM	95.05	95.22	95.05	95.05	**98.65 ± 1.44**	**98.74 ± 1.26**	**98.67 ± 1.33**	**99.06 ± 0.04**
	B-CS-AM-NAM-D	96.87	97.04	96.87	96.87	97.91 ± 2.08	97.80 ± 2.10	97.82 ± 2.18	97.56 ± 2.40
	B-CS-D	**97.40**	**97.49**	**97.40**	**97.39**	98.28 ± 1.70	98.04 ± 1.66	97.30 ± 2.29	97.24 ± 2.65
	B-CS-NAM-D	97.13	97.29	97.13	97.13	96.79 ± 3.03	97.86 ± 2.07	97.60 ± 2.06	97.43 ± 2.36
	B-CS-NAM	95.31	95.35	95.31	95.31	96.77 ± 3.23	96.99 ± 3.01	96.95 ± 3.05	97.03 ± 2.97
LR	B-CS	99.74	99.74	99.74	99.74	99.12 ± 0.88	99.13 ± 0.87	99.12 ± 0.88	99.12 ± 0.88
	B-CS-AM	98.96	98.98	98.96	98.96	99.27 ± 0.73	99.29 ± 0.71	99.27 ± 0.73	99.27 ± 0.73
	B-CS-AM-D	98.96	98.98	98.96	98.96	99.30 ± 0.70	99.30 ± 0.70	99.30 ± 0.70	99.30 ± 0.70
	B-CS-AM-NAM	99.74	99.74	99.74	99.74	99.30 ± 0.70	99.32 ± 0.68	99.30 ± 0.70	99.30 ± 0.70
	B-CS-AM-NAM-D	**99.74**	**99.74**	**99.74**	**99.74**	**99.30 ± 0.70**	**99.32 ± 0.68**	**99.30 ± 0.68**	**99.30 ± 0.68**
	B-CS-D	99.22	99.22	99.24	99.22	99.15 ± 0.85	99.18 ± 0.82	99.15 ± 0.85	99.15 ± 0.85
	B-CS-NAM-D	99.22	99.22	99.24	99.22	99.27 ± 0.73	99.28 ± 0.30	99.27 ± 0.73	99.27 ± 0.73
	B-CS-NAM	99.48	99.48	99.48	99.48	99.25 ± 0.75	99.24 ± 0.76	99.23 ± 0.77	99.25 ± 0.75
SVM	B-CS	100.0	100.0	100.0	100.0	98.69 ± 1.30	98.71 ± 1.29	98.69 ± 1.30	98.69 ± 1.30
	B-CS-AM	100.0	100.0	100.0	100.0	99.15 ± 0.85	99.16 ± 0.84	99.15 ± 0.85	99.15 ± 0.85
	B-CS-AM-D	94.79	95.28	94.79	94.77	99.44 ± 0.56	99.46 ± 0.54	99.44 ± 0.56	99.44 ± 0.56
	B-CS-AM-NAM	**100.0**	**100.0**	**100.0**	**100.0**	99.33 ± 0.67	99.35 ± 0.65	99.33 ± 0.67	99.33 ± 0.67
	B-CS-AM-NAM-D	96.61	96.83	96.61	96.61	**99.59 ± 0.41**	**99.61 ± 0.39**	**99.59 ± 0.41**	**99.59 ± 0.41**
	B-CS-D	100.0	100.0	100.0	100.0	99.03 ± 0.97	99.06 ± 0.94	99.03 ± 0.97	99.03 ± 0.97
	B-CS-NAM-D	100.0	100.0	100.0	100.0	99.27 ± 0.70	99.29 ± 0.70	99.27 ± 0.70	99.27 ± 0.70
	B-CS-NAM	100.0	100.0	100.0	100.0	99.40 ± 0.60	99.41 ± 0.59	99.40 ± 0.60	99.40 ± 0.60
KNN	B-CS	98.44	98.48	98.44	98.44	98.95 ± 1.05	98.99 ± 1.04	98.95 ± 1.05	98.95 ± 1.05
	B-CS-AM	98.44	98.48	98.44	98.44	98.98 ± 1.02	99.02 ± 0.08	98.98 ± 1.02	98.98 ± 1.02
	B-CS-AM-D	**98.44**	**98.48**	**98.44**	**98.44**	**99.06 ± 0.94**	**99.10 ± 0.90**	**99.06 ± 0.94**	**99.06 ± 0.94**
	B-CS-AM-NAM	98.44	98.48	98.44	98.44	98.94 ± 1.04	98.98 ± 1.02	98.94 ± 1.04	98.94 ± 1.04
	B-CS-AM-NAM-D	98.44	98.48	98.44	98.44	98.94 ± 1.04	98.98 ± 1.02	98.94 ± 1.04	98.94 ± 1.04
	B-CS-D	98.44	98.48	98.44	98.44	98.91 ± 1.09	98.96 ± 1.04	98.91 ± 1.09	98.91 ± 1.09
	B-CS-NAM-D	98.44	98.48	98.44	98.44	98.89 ± 1.11	98.92 ± 1.18	98.89 ± 1.11	98.89 ± 1.11
	B-CS-NAM	98.44	98.48	98.44	98.44	98.89 ± 1.11	98.93 ± 1.17	98.89 ± 1.11	98.89 ± 1.11

to detect pMCI patients. In this subsection, we will discuss the cross-validation results.

DT achieved the highest performances by using the B-CS-NAM dataset (accuracy of 98.44%, precision of 98.19%, recall of 98.29%, and F1-score of 97.84%). It achieved high improvements compared to the B-CS data (accuracy of +13.5%, precision of +14.7%, recall of +12.7%, and F1-score of +14.3%). By using the B-CS-NAM-D dataset, DT showed the lowest performances (accuracy of 78.48%, precision of 80.01%, recall of 79.52%, and F1-score of 79.68%). RF saw slight improvements (+3%) using the B-CS-AM-NAM dataset compared to baseline (accuracy of 94%, precision of 95.1%, recall of 94.91%, and F1-score of 94.96%). LR achieved good performance improvements using all fused datasets. For accuracy, LR has gained +1.0%, +2.1%, +9.3%, +4.4%, +1.9%, +2.8%, and +2.8% with the B-CS-AM, B-CS-AM-D, B-CS-AM-NAM, B-CS-AM-NAM-D, B-CS-D, B-CS-NAM-D, and B-CS-NAM datasets, respectively.

LR achieved the highest improvement using the B-CS-AM-NAM dataset (precision of +8.9%, recall of +9.3%, and F1-score of +9.3%). Similarly, KNN gained significant performance improvement using the B-CS-AM-NAM dataset (accuracy of 93.7%, precision of 93.86%, recall of 93.71%, and F1-score of 93.7%). The improvement with respect to the baseline is +9.77% of accuracy, +8.43% for precision, +9.77% for recall, and +9.9% for F1-score. Furthermore, KNN gained non-negligible improvement by using the B-CS-NAM-D and B-CS-NAM datasets (+0.6%, +1.4%, +0.6%, and +0.5 for accuracy, precision, recall, and F1-score, respectively). However, SVM did not gain any performance improvements for all fused dataset.

3.4.6. Testing results: AD vs. pMCI

The performance of the ML models on unseen data for the AD vs. pMCI classification is presented in Table 6. RF and SVM showed the best performance for all fused datasets. For RF, the highest performance is obtained by using B-CS-AM-D and B-CS-AM-NAM datasets, the improvements compared to the B-CS dataset were accuracy of +4.6%, precision of +4.6%, recall of +4.7% and F1-score of +4.6% for B-CS-AM-D dataset; and accuracy of +4.6%, precision of +4.6%, recall of +4.6%, and F1-score of +4.6% for B-CS-AM-NAM dataset. LR obtains the highest performance improvement using the B-CS-AM-NAM dataset (accuracy of +4.8%, precision of +4.6%, recall of +4.8%, and F1-score of +4.8%). On the other hand, DT showed lower improvements by using fused sets except for the B-CS-D dataset (accuracy of +0.6%, recall of +0.6%, and F1-score of +0.8%). KNN achieved higher performance using B-CS-AM, B-CS-AM-NAM-D, and B-CS-NAM-D. For example, KNN showed a performance improvement of +0.3% for accuracy, +0.3% for precision, +0.2% for recall, and +0.8% for F1-score by using the B-CS-AM dataset. SVM also showed improvement using all fused datasets expect B-CS-AM-NAM-D. It showed the highest improvement by using the B-CS-NAM-D dataset (accuracy of +5%, precision of +4%, recall of +5%, and F1-score of +5%).

Table 5
Performance for the AD vs. sMCI task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	90.74	90.79	90.74	90.74	93.38 ± 2.80	93.42 ± 2.70	93.33 ± 2.82	93.35 ± 2.82
	B-CS-AM	**92.36**	**92.40**	**92.36**	**92.35**	94.68 ± 2.17	94.81 ± 2.05	94.71 ± 2.06	94.60 ± 2.13
	B-CS-AM-D	91.43	91.57	91.43	91.43	94.68 ± 2.16	94.63 ± 2.26	94.71 ± 2.19	94.50 ± 2.34
	B-CS-AM-NAM	92.13	92.19	92.13	92.12	**94.91 ± 2.41**	**95.10 ± 2.38**	**94.91 ± 2.41**	**94.96 ± 2.36**
	B-CS-AM-NAM-D	91.43	91.56	91.43	91.42	94.73 ± 2.62	94.75 ± 2.50	94.65 ± 2.71	94.78 ± 2.30
	B-CS-D	91.43	91.49	91.43	91.43	93.38 ± 2.62	93.57 ± 2.62	93.61 ± 2.63	93.60 ± 2.74
	B-CS-NAM-D	89.81	89.85	89.81	89.81	94.12 ± 2.06	94.36 ± 1.92	94.15 ± 1.98	94.12 ± 2.12
	B-CS-NAM	90.51	90.57	90.51	90.50	94.12 ± 2.15	94.30 ± 2.07	94.12 ± 1.96	94.30 ± 2.35
DT	B-CS	85.87	86.02	85.87	85.86	**91.70 ± 3.26**	**91.94 ± 3.60**	**91.54 ± 3.38**	**90.83 ± 3.15**
	B-CS-AM	83.79	84.18	83.79	83.74	89.96 ± 3.44	91.63 ± 3.66	90.19 ± 3.46	89.73 ± 3.90
	B-CS-AM-D	84.49	84.91	84.49	84.43	89.68 ± 3.92	89.12 ± 4.22	89.05 ± 4.65	88.95 ± 4.48
	B-CS-AM-NAM	86.11	86.24	86.11	86.10	89.36 ± 4.22	90.09 ± 3.38	89.67 ± 3.63	90.09 ± 3.45
	B-CS-AM-NAM-D	82.17	82.41	82.17	82.14	88.55 ± 3.84	89.33 ± 4.52	89.37 ± 4.11	88.67 ± 4.59
	B-CS-D	**88.65**	**88.76**	**88.65**	**88.65**	89.24 ± 3.81	89.59 ± 3.01	89.09 ± 4.26	89.15 ± 3.93
	B-CS-NAM-D	81.01	81.15	81.01	80.99	88.67 ± 3.85	88.78 ± 3.82	87.92 ± 3.68	88.69 ± 3.81
	B-CS-NAM	84.25	84.57	84.25	84.22	90.85 ± 2.91	90.88 ± 2.84	90.03 ± 3.18	90.47 ± 2.77
LR	B-CS	86.57	86.65	86.57	86.56	92.65 ± 3.27	92.84 ± 3.20	92.65 ± 3.27	92.64 ± 3.28
	B-CS-AM	88.19	88.29	88.19	88.18	92.90 ± 3.16	93.06 ± 3.10	92.90 ± 3.16	92.89 ± 3.17
	B-CS-AM-D	**89.12**	**89.12**	**89.12**	**89.12**	92.46 ± 3.35	92.62 ± 3.32	92.46 ± 3.35	92.46 ± 3.36
	B-CS-AM-NAM	87.50	87.54	87.50	87.49	**93.04 ± 2.92**	**93.20 ± 2.88**	**93.04 ± 2.92**	**93.03 ± 2.92**
	B-CS-AM-NAM-D	85.64	85.84	85.64	85.62	92.71 ± 2.96	92.87 ± 2.94	92.71 ± 2.96	92.70 ± 2.97
	B-CS-D	88.89	88.89	88.89	88.89	91.66 ± 2.69	91.86 ± 2.63	91.66 ± 2.69	91.65 ± 2.70
	B-CS-NAM-D	89.12	89.12	89.12	89.12	91.89 ± 2.50	92.11 ± 2.47	91.90 ± 2.47	91.86 ± 2.51
	B-CS-NAM	88.65	88.68	88.65	88.65	92.53 ± 2.24	92.69 ± 2.26	92.53 ± 2.24	92.52 ± 2.24
SVM	B-CS	83.10	83.25	83.10	83.08	93.46 ± 2.82	93.63 ± 2.73	93.46 ± 2.82	93.45 ± 2.83
	B-CS-AM	83.56	83.61	83.56	83.55	**93.97 ± 2.72**	**94.09 ± 2.70**	**93.97 ± 2.72**	**93.96 ± 2.72**
	B-CS-AM-D	84.95	84.97	84.95	84.95	92.03 ± 3.29	92.21 ± 3.25	92.03 ± 3.29	92.02 ± 3.29
	B-CS-AM-NAM	86.80	87.55	86.80	86.74	91.97 ± 3.32	92.22 ± 3.24	91.97 ± 3.32	91.95 ± 3.33
	B-CS-AM-NAM-D	**90.51**	**90.69**	**90.51**	**90.49**	92.28 ± 3.07	92.50 ± 3.00	92.28 ± 3.07	92.26 ± 3.07
	B-CS-D	88.65	88.76	88.65	88.65	92.77 ± 2.51	92.96 ± 2.45	92.77 ± 2.51	92.76 ± 2.52
	B-CS-NAM-D	88.89	89.31	88.89	88.85	93.08 ± 2.30	93.27 ± 2.24	93.08 ± 2.30	93.07 ± 2.31
	B-CS-NAM	86.57	86.82	86.57	86.55	92.68 ± 2.16	92.85 ± 2.14	92.68 ± 2.16	92.67 ± 2.16
KNN	B-CS	90.51	90.63	90.51	90.50	93.71 ± 2.33	93.88 ± 2.27	93.71 ± 2.33	93.70 ± 2.34
	B-CS-AM	90.51	90.63	90.51	90.50	**93.74 ± 2.40**	**93.91 ± 2.32**	**93.74 ± 2.40**	**93.73 ± 2.40**
	B-CS-AM-D	91.90	91.95	91.90	91.89	93.56 ± 2.41	93.73 ± 2.35	93.56 ± 2.41	93.55 ± 2.42
	B-CS-AM-NAM	91.90	91.95	91.90	91.89	93.71 ± 3.18	93.86 ± 3.15	93.71 ± 3.18	93.71 ± 3.19
	B-CS-AM-NAM-D	90.74	90.80	90.74	90.74	93.61 ± 3.06	93.76 ± 3.01	93.61 ± 3.06	93.61 ± 3.06
	B-CS-D	91.20	91.30	91.20	91.20	92.46 ± 2.49	92.81 ± 2.26	92.46 ± 2.49	92.44 ± 2.52
	B-CS-NAM-D	**92.82**	**92.87**	**92.82**	**92.81**	93.41 ± 2.03	93.63 ± 1.92	93.41 ± 2.03	93.40 ± 2.04
	B-CS-NAM	91.20	91.34	91.20	91.19	93.41 ± 2.16	93.64 ± 2.02	93.41 ± 2.16	93.40 ± 2.17

3.4.7. Results for the AD vs. pMCI task

The CV performance of the ML models for the CN vs. MCI classification are presented in Table 7. RF achieved the best accuracy of 95.67%±1.62 based on the B-CS-D data, and the best testing performance (i.e. accuracy of 95.60%) using the B-CS-AM data. LR and SVM achieved comparable CV performance to RF based on the B-CS-NAM data. The achieved CV accuracies were 95.09%±2.63 and 95.25%±2.46 and the testing accuracies were 87.50% and 89.12%, respectively. DT achieved a CV accuracy of 91.95%±3.64 based on the B-CS data, and testing accuracy of 86.57% based on the B-CS-NAM-D data. KNN achieved the lowest CV accuracy of 80.43%±3.95 based on the B-CS-NAM-D data, and the lowest testing accuracy of 81.94% based on the B-CS-AM-NAM data.

Fig. 14 illustrates the performance of the best models from the five different experiments based on the accuracy metric. For the 4-class and 3-class tasks, RF achieves the best accuracies of 90.51% and 91.21% by using the B-CS-AM-D and B-CS-AM-NAM dataset, respectively. SVM showed the best accuracy (93.23%) for the AD vs. pMCI task by utilizing the B-CS-NAM-D dataset. For the AD vs. sMCI task, KNN achieves the best accuracy (92.82%) by using the B-CS-NAM-D dataset. RF is the best model for the AD vs. CN task (accuracy of 100%). Finally, RF achieved the best accuracy (95.60%) for the CN vs. MCI task based on the B-CS-AM data. Although the 4-class task is medically a challenge, using our fused data, the ML models achieved promising results. In addition, the performance was improved by reducing the problem

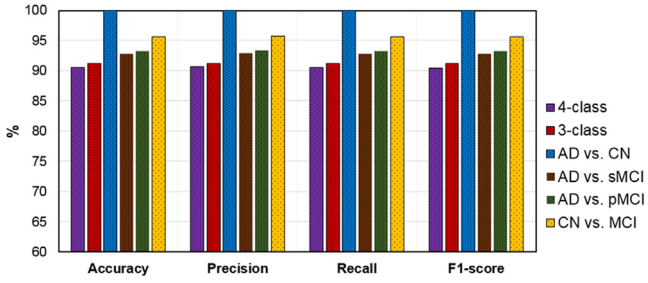

Fig. 14. Comparison among different experiments.

to 3-class and 2-class tasks. RF achieved the best results in most of the experiments, and it is the most stable technique. The fused dataset of B-CS-NAM-D is utilized by RF, KNN, and SVM to achieve the best binary classification results. The comorbidities features are replaced with brain disease medications (i.e. B-CS-AM-NAM) for the 3-class task. Therefore, these results obviously highlight the crucial role of the newly utilized medications and comorbidities data.

For the 4-class task, the B-CS-AM-NAM dataset achieved the best CV performance, B-CS-AM-D achieved the best testing performance. For the 3-class task, the same dataset (i.e. B-CS-AM-NAM) achieved the best performance. For the binary classification

Table 6
Performance for the AD vs. pMCI task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	86.46	86.97	86.46	86.41	92.04 ± 3.58	92.09 ± 3.33	91.92 ± 3.74	92.25 ± 3.36
	B-CS-AM	89.58	90.04	89.58	89.55	88.92 ± 3.08	89.51 ± 3.10	89.07 ± 3.02	89.02 ± 3.17
	B-CS-AM-D	90.62	91.27	90.62	90.59	90.18 ± 2.72	90.82 ± 2.73	90.41 ± 2.96	90.21 ± 3.01
	B-CS-AM-NAM	**90.62**	**91.02**	**90.62**	**90.60**	**94.90 ± 2.41**	**95.10 ± 2.38**	**94.91 ± 2.41**	**94.96 ± 2.36**
	B-CS-AM-NAM-D	90.36	90.71	90.36	90.34	91.25 ± 3.68	91.30 ± 3.58	90.92 ± 3.67	91.10 ± 3.95
	B-CS-D	86.45	87.05	86.45	86.40	89.03 ± 3.32	89.73 ± 3.26	89.36 ± 3.70	88.87 ± 3.42
	B-CS-NAM-D	86.72	87.12	86.72	86.68	89.61 ± 3.83	90.08 ± 3.44	89.82 ± 3.57	89.41 ± 3.69
	B-CS-NAM	86.98	87.25	86.98	86.95	90.27 ± 3.59	91.09 ± 3.25	90.50 ± 3.71	90.11 ± 3.73
DT	B-CS	80.99	82.21	80.99	80.81	85.12 ± 4.75	83.71 ± 4.98	85.79 ± 3.72	83.81 ± 4.54
	B-CS-AM	78.64	78.98	78.64	78.57	84.70 ± 3.80	85.71 ± 4.07	84.65 ± 4.16	85.1 ± 4.30
	B-CS-AM-D	76.82	76.96	76.82	76.74	83.50 ± 4.91	82.30 ± 4.51	81.94 ± 5.14	80.94 ± 5.66
	B-CS-AM-NAM	78.12	78.24	78.12	78.10	89.36 ± 4.22	90.09 ± 3.38	89.67 ± 3.63	90.09 ± 3.45
	B-CS-AM-NAM-D	79.94	80.88	79.94	79.78	82.23 ± 5.25	81.98 ± 5.22	81.69 ± 6.33	81.14 ± 6.15
	B-CS-D	**81.51**	**81.62**	**81.51**	**81.49**	82.76 ± 4.99	82.20 ± 4.94	82.35 ± 5.30	81.88 ± 5.01
	B-CS-NAM-D	73.17	74.04	73.17	72.92	78.48 ± 7.05	80.01 ± 6.01	79.52 ± 6.08	79.68 ± 5.16
	B-CS-NAM	76.30	76.49	76.30	76.25	**98.44 ± 1.46**	**98.19 ± 1.65**	**98.29 ± 1.62**	**97.84 ± 2.12**
LR	B-CS	83.59	84.65	83.59	83.47	84.41 ± 4.29	84.90 ± 4.25	84.38 ± 4.34	84.35 ± 4.32
	B-CS-AM	86.46	88.13	86.46	86.46	85.30 ± 3.92	85.65 ± 3.87	85.24 ± 3.90	85.23 ± 3.96
	B-CS-AM-D	87.24	88.13	87.24	87.16	86.18 ± 3.61	86.56 ± 3.60	86.13±.615	86.11 ± 3.61
	B-CS-AM-NAM	**87.76**	**88.70**	**87.76**	**87.68**	**93.04 ± 2.93**	**93.20 ± 2.88**	**93.04 ± 2.92**	**93.04 ± 2.92**
	B-CS-AM-NAM-D	86.98	87.68	86.98	86.91	88.27 ± 3.97	88.68 ± 3.88	88.27 ± 3.97	88.23 ± 3.99
	B-CS-D	86.19	86.55	86.19	86.16	86.03 ± 4.59	86.64 ± 4.44	86.03 ± 4.59	85.96 ± 4.63
	B-CS-NAM-D	84.37	85.03	84.37	84.30	86.86 ± 4.27	87.30 ± 4.22	86.83 ± 4.23	86.82 ± 4.29
	B-CS-NAM	84.89	85.30	84.89	84.85	86.86 ± 4.27	87.30 ± 4.22	86.83 ± 4.23	86.82 ± 4.29
SVM	B-CS	88.80	89.19	88.80	88.77	**93.70 ± 2.84**	**93.91 ± 2.76**	**93.70 ± 2.84**	**93.69 ± 2.85**
	B-CS-AM	90.10	90.32	90.10	90.09	92.42 ± 3.08	92.77 ± 2.93	92.42 ± 3.08	92.40 ± 3.09
	B-CS-AM-D	90.36	90.41	90.36	90.35	91.31 ± 3.18	91.96 ± 3.04	91.31 ± 3.18	91.28 ± 3.21
	B-CS-AM-NAM	89.84	89.87	89.84	89.84	90.69 ± 3.17	91.23 ± 3.01	90.69 ± 3.17	90.65 ± 3.20
	B-CS-AM-NAM-D	88.02	88.05	88.02	88.01	90.63 ± 3.31	91.21 ± 3.22	90.63 ± 3.31	90.59 ± 3.33
	B-CS-D	92.96	93.08	92.96	92.96	89.58 ± 3.49	90.06 ± 3.36	89.58 ± 3.49	89.54 ± 3.52
	B-CS-NAM-D	**93.23**	**93.35**	**93.23**	**93.22**	89.30 ± 3.48	89.97 ± 3.20	89.30 ± 3.48	89.24 ± 3.53
	B-CS-NAM	92.19	92.19	92.19	92.19	89.30 ± 3.48	89.97 ± 3.20	89.30 ± 3.48	89.24 ± 3.53
KNN	B-CS	78.12	82.73	78.12	77.32	84.55 ± 3.42	85.95 ± 3.25	84.55 ± 3.42	84.39 ± 3.50
	B-CS-AM	**78.38**	**82.88**	**78.38**	**77.62**	84.52 ± 3.54	85.96 ± 3.39	84.52 ± 3.54	84.35 ± 3.64
	B-CS-AM-D	78.12	82.73	78.12	77.32	84.52 ± 3.52	86.05 ± 3.39	84.52 ± 3.52	84.34 ± 3.61
	B-CS-AM-NAM	77.08	82.11	77.08	76.14	**93.71 ± 3.18**	**93.86 ± 3.15**	**93.71 ± 3.18**	**93.70 ± 3.19**
	B-CS-AM-NAM-D	77.34	84.03	77.34	76.17	83.91 ± 3.65	86.57 ± 3.12	83.91 ± 3.65	83.58 ± 3.84
	B-CS-D	76.82	81.96	76.82	75.84	85.40 ± 3.31	87.67 ± 2.81	85.40 ± 3.31	85.15 ± 3.43
	B-CS-NAM-D	77.08	84.28	77.08	75.81	85.06 ± 3.52	87.17 ± 2.98	85.06 ± 3.52	84.81 ± 3.66
	B-CS-NAM	76.82	81.96	76.82	75.84	85.06 ± 3.52	87.17 ± 2.98	85.06 ± 3.52	84.81 ± 3.66

tasks, B-CS-NAM-D achieved the best results for the AD vs. CN task, B-CS-AM-NAM achieved the best results for the AD vs. sMCI task, and B-CS-AM-NAM achieved the best results for the AD vs. pMCI task.

As can be seen from these results, the Alzheimer's medicine data (AM) provided critical and discriminative features in most tasks. In addition, non-Alzheimer's medication (NAM) had a critical role in the 4-class, 3-class, and all binary tasks. As Alzheimer's is a complex chronic disease, patients always have many comorbidities such as hypertension and diabetes. These comorbidity features helped all classifiers to achieve the best results in the 4-class task. Medically, the taken drugs for these diseases could affect the speed of Alzheimer's progression. Unfortunately, there are no studies that explore this issue in the literature. On the other hand, large and longitudinal datasets like ADNI have collected these features. In our study, we investigated the role of medication and comorbidity modalities and discovered a big relationship between these features and the accuracy of disease progression detection classifiers. These features have been studied in many experiments with different complexities (i.e. 4-class, 3-class, and 2-class experiments). To study the role of different individual features in each experiment, we calculated the feature importance for each case by using two popular techniques: permutation importance and SHAP [40]. The importance of the first 20 features for every case is shown in Fig. 15.

Permutation importance is more accurate than the feature importance calculated by random forest classifiers. The RF-based method was computed using the statistics derived from training data. The highest ranks could be linked to features that are not predictive of the target variable. SHAP values are calculated based on the shapely values from game theory. The calculated ranks for every task are consistent between the Permutation importance and SHAP techniques. As shown in Fig. 15, the cognitive scores have definitely the highest ranks for all cases and using the two techniques. Please note that cognitive scores are also more important than any other modalities including neuroimaging and lab tests [23]. However, we can find other features from drugs and comorbidities that have higher ranks compared to the demographics and medical history features. we concentrate on the comorbidity and drug modalities in this discussion. For example, in the 4-class task, the Aricept drug has higher importance than the geriatric depression scale, patient education, and even gender. Further, Namenda has a high rank compared to other features, and knowing that the patient is not taking any Alzheimer's drugs is more important than patient age. Please note that we selected only the top 20 features out of 54 features. The same behavior was observed for the other tasks. We observe that the drug groups of C, H, M, A, and R have important roles in these classification tasks. As clearly reported in all binary classification tasks, the comorbidity modality can play a great role in improving model performance. This observation is asserted in the calculated feature importance for all binary tasks. For example, MHPSYCH, MH13ALLE, MH16SMOK, and MHPSYCH all occupy a high rank.

Table 7
Performance for the CN vs. MCI task.

Model	Dataset	Testing performance				Cross-validation performance			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
RF	B-CS	92.82	93.07	92.82	92.81	94.90 ± 2.90	95.01 ± 2.82	95.01 ± 2.83	94.88 ± 2.83
	B-CS-AM	**95.60**	**95.69**	**95.60**	**95.59**	95.31 ± 2.54	95.47 ± 2.39	95.36 ± 2.37	95.28 ± 2.51
	B-CS-AM-D	93.05	93.27	93.05	93.04	95.19 ± 2.69	95.35 ± 2.59	95.06 ± 2.73	95.05 ± 2.71
	B-CS-AM-NAM	93.28	93.43	93.28	93.28	95.16 ± 2.56	95.32 ± 2.48	95.21 ± 2.45	95.26 ± 2.55
	B-CS-AM-NAM-D	93.28	93.70	93.28	93.27	95.47 ± 2.52	95.38 ± 2.58	95.26 ± 2.54	95.26 ± 2.51
	B-CS-D	92.59	92.61	92.59	92.59	**95.67 ± 1.62**	**95.69 ± 1.64**	**95.74 ± 1.71**	**95.55 ± 1.63**
	B-CS-NAM-D	93.98	94.06	93.98	93.97	95.59 ± 2.46	95.58 ± 2.32	95.63 ± 2.26	95.49 ± 2.42
	B-CS-NAM	94.67	94.71	94.67	94.67	95.63 ± 2.57	95.69 ± 2.53	95.61 ± 2.55	95.57 ± 2.50
DT	B-CS	84.48	84.83	84.48	84.45	**91.95 ± 3.64**	**91.18 ± 4.16**	**91.72 ± 4.27**	**91.83 ± 3.91**
	B-CS-AM	84.02	84.37	84.02	83.98	88.81 ± 5.88	90.65 ± 5.63	90.00 ± 6.03	90.41 ± 4.66
	B-CS-AM-D	78.93	79.37	78.93	78.83	86.17 ± 7.05	87.31 ± 7.37	82.87 ± 6.83	85.79 ± 7.20
	B-CS-AM-NAM	86.11	86.20	86.11	86.10	89.66 ± 5.94	89.82 ± 5.39	88.79 ± 6.21	86.81 ± 6.41
	B-CS-AM-NAM-D	80.55	81.05	80.55	80.45	84.70 ± 7.66	86.51 ± 5.97	86.38 ± 6.26	87.14 ± 6.93
	B-CS-D	84.95	85.22	84.95	84.91	88.76 ± 5.39	87.89 ± 4.43	88.69 ± 4.88	89.06 ± 4.70
	B-CS-NAM-D	**86.57**	**86.64**	**86.57**	**86.56**	86.20 ± 5.65	88.34 ± 5.48	87.95 ± 5.71	87.81 ± 5.37
	B-CS-NAM	83.10	83.39	83.10	83.04	87.98 ± 6.36	87.72 ± 5.61	89.88 ± 6.03	87.89 ± 6.29
LR	B-CS	87.26	87.29	87.26	87.26	95.09 ± 2.68	95.22 ± 2.61	95.09 ± 2.68	95.08 ± 2.69
	B-CS-AM	88.65	88.69	88.65	88.65	94.88 ± 2.78	95.03 ± 2.68	94.88 ± 2.78	94.87 ± 2.78
	B-CS-AM-D	88.19	88.22	88.19	88.19	94.88 ± 2.79	95.03 ± 2.69	94.88 ± 2.79	94.87 ± 2.80
	B-CS-AM-NAM	88.19	88.33	88.19	88.18	94.78 ± 2.79	94.91 ± 2.72	94.78 ± 2.79	94.77 ± 2.80
	B-CS-AM-NAM-D	87.73	87.82	87.73	87.72	93.57 ± 3.05	93.72 ± 3.01	93.57 ± 3.05	93.57 ± 3.05
	B-CS-D	88.42	88.47	88.42	88.42	94.67 ± 1.89	94.78 ± 1.88	94.67 ± 1.89	94.67 ± 1.89
	B-CS-NAM-D	87.50	87.55	87.50	87.49	94.34 ± 2.50	94.42 ± 2.46	94.34 ± 2.50	94.34 ± 2.51
	B-CS-NAM	**87.50**	**87.73**	**87.50**	**87.48**	**95.09 ± 2.63**	**95.16 ± 2.61**	**95.09 ± 2.63**	**95.09 ± 2.63**
SVM	B-CS	85.87	85.96	85.87	85.87	94.70 ± 2.90	94.81 ± 2.85	94.70 ± 2.90	94.70 ± 2.91
	B-CS-AM	87.03	87.09	87.03	87.03	94.98 ± 2.72	95.10 ± 2.67	94.98 ± 2.72	94.98 ± 2.73
	B-CS-AM-D	85.41	85.56	85.41	85.40	94.65 ± 2.61	94.79 ± 2.52	94.65 ± 2.61	94.64 ± 2.62
	B-CS-AM-NAM	88.19	88.35	88.19	88.18	95.06 ± 2.23	95.19 ± 2.16	95.06 ± 2.23	95.06 ± 2.23
	B-CS-AM-NAM-D	87.26	87.43	87.26	87.25	93.48 ± 2.46	93.68 ± 2.37	93.48 ± 2.46	93.47 ± 2.47
	B-CS-D	83.10	83.14	83.10	83.09	95.04 ± 1.96	95.15 ± 1.92	95.04 ± 1.96	95.03 ± 1.96
	B-CS-NAM-D	83.79	83.89	83.79	83.78	94.38 ± 2.18	94.45 ± 2.19	94.38 ± 2.18	94.38 ± 2.18
	B-CS-NAM	**89.12**	**89.32**	**89.12**	**89.10**	**95.25 ± 2.46**	**95.32 ± 2.46**	**95.25 ± 2.46**	**95.24 ± 2.46**
KNN	B-CS	78.01	81.20	78.01	77.43	79.58 ± 3.83	82.12 ± 3.69	79.58 ± 3.83	79.14 ± 4.05
	B-CS-AM	78.01	81.20	78.01	77.43	79.32 ± 3.97	82.08 ± 3.72	79.32 ± 3.97	78.84 ± 4.24
	B-CS-AM-D	81.25	83.58	81.25	80.90	80.37 ± 3.96	82.73 ± 3.89	80.37 ± 3.96	79.99 ± 4.14
	B-CS-AM-NAM	**81.94**	**83.91**	**81.94**	**81.67**	79.90 ± 4.83	81.70 ± 4.59	79.90 ± 4.83	79.58 ± 5.01
	B-CS-AM-NAM-D	78.47	81.92	78.47	77.87	79.30 ± 5.03	81.68 ± 4.82	79.32 ± 5.02	78.88 ± 5.23
	B-CS-D	79.86	82.84	79.86	79.39	80.07 ± 3.57	81.67 ± 3.41	80.07 ± 3.57	79.80 ± 3.69
	B-CS-NAM-D	80.32	83.80	80.32	79.80	**80.43 ± 3.95**	**81.94 ± 3.82**	**80.43 ± 3.95**	**80.19 ± 4.06**
	B-CS-NAM	78.70	82.54	78.70	78.05	79.65 ± 3.88	81.21 ± 3.64	79.65 ± 3.88	79.37 ± 4.02

These results assert that cost-effective features like drugs taken and patient history can be used to predict the future status of a patient with high accuracy.

Most previous studies modeled AD progression as binary classification problems (e.g. CN vs. AD [41], MCI vs. AD [18], sMCI vs. pMCI [42,43]). Westman et al. [44] utilized the MRI and CSF data to achieve an accuracy of 91.8% for classifying AD vs. CN, this performance dropped to 71.8% for MCI vs. CN. Chincarini et al. [24] utilized longitudinal hippocampal volume features (i.e. four time-steps) and achieved an AUC of 0.93 for CN vs. AD and an AUC of 0.88 for CN vs. MCI. Tangaro et al. [45] utilized baseline MRIs and cognitive measurement data of 372 ADNI patients to predict AD progression using SVM. From MRI features, the study concentrated on the role of hippocampal volume to predict AD progression after one year. The model handled the uncertainty degrees that affect neuroimaging features by using fuzzy logic. The authors trained the model using CN vs. AD classes and tested it using the sMCI vs. pMCI task. This model achieved an AUC of 88.2%. On the other hand, some studies have modeled AD progression as a 4-class (CN vs. sMCI vs. pMCI vs. AD) classification task [44,46]. All of these studies achieved lower performance than ours. For example, Yao et al. [47] used a hierarchical ensemble and baseline data of MRI, age, gender, and MMSE to achieve 54.38% accuracy. Amoroso et al. [48] achieved an accuracy of 56.3% using a deep neural network classifier with RF for the feature selection step. Nanni et al. [49] achieved 52.92% accuracy using a voting classifier and baseline data of MRI, age, and MMSE. Liu et al. [50] achieved an accuracy of 51.8% based on a CNN and baseline MRI data. Sorensen et al. [51] achieved 59.10% accuracy using bagging and baseline data of MRI, age, gender, and MMSE. Other studies achieved similar accuracies such as Dimitriadis and Liparas [52] (61.90%), Ramírez et al. [53] (56.25%), and Jin and Deng [A2](56.25%). Relaxing the 4-class problem to a 3-class one enhances the overall performance. For example, Moore et al. [46] achieved 73% accuracy based on an RF classifier and baseline data of MRI and CSs. In the CADDementia challenge [12], the best algorithm achieved an accuracy of 63.0% and, an AUC of 78.8% for AD diagnosis as a three class classification task (i.e. CN vs. MCI vs. AD). Notice that all the best models are based on fused datasets. To the best of our knowledge, there are no studies in the literature which discuss the same issue or use similar data to ours. The resulting models achieved higher results than most state-of-the-art models in current literature [47,52–54]. Furthermore, the proposed models are cost-effective and can be easily implemented in healthcare environments. Besides, by using these models, patients can check at home their current situation without the need for any neuroimaging scans.

4. Conclusion

In this paper, we investigated the role of a new cost-effective and easy to collect set of features for the prediction of AD progression. We prepared a set of three time-series modalities consisting

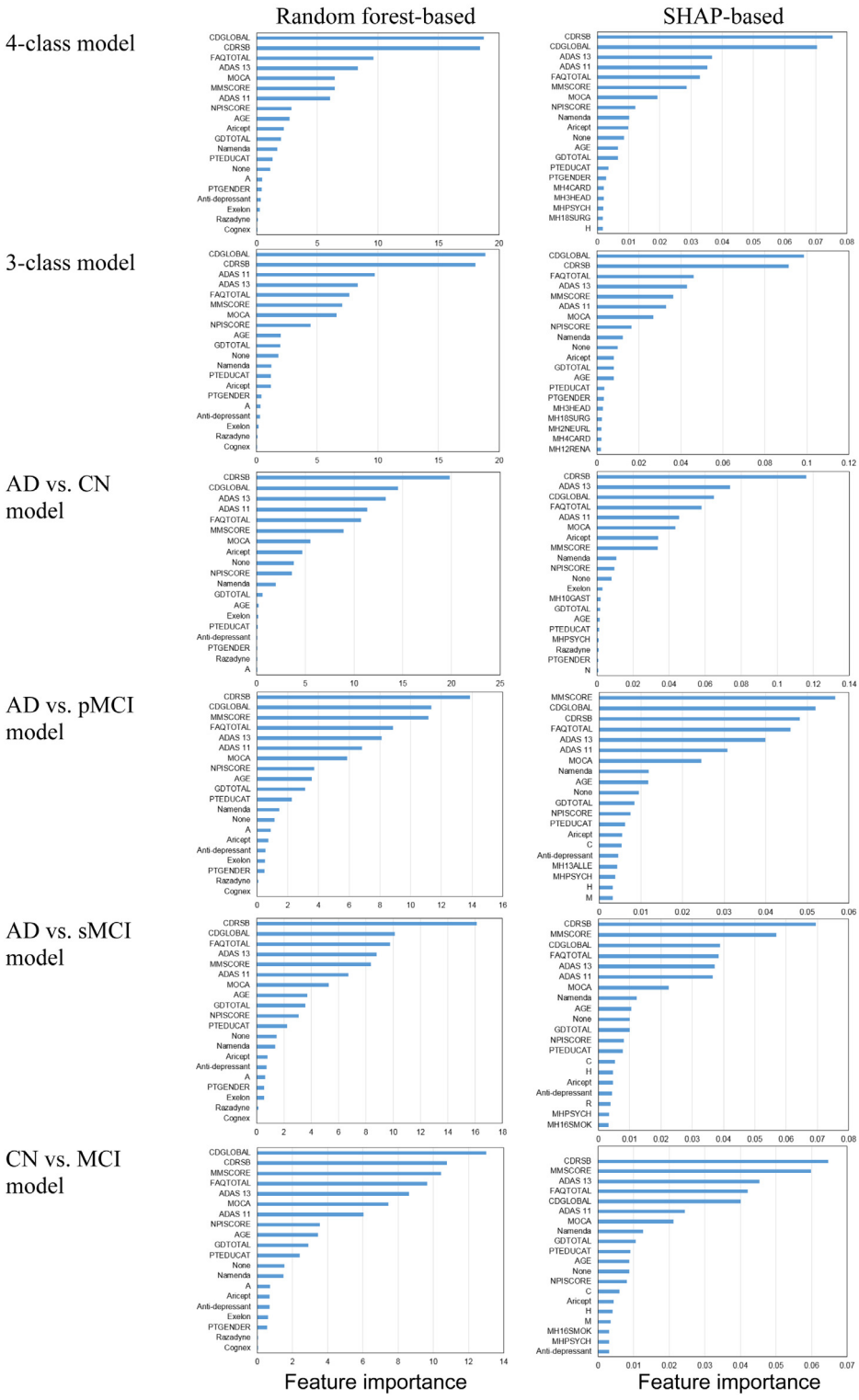

Fig. 15. Feature importance based on RF and SHAP for all cases.

of cognitive scores, medications, and comorbidities. These data were grouped into four time-steps (i.e., baseline, M06, M12, and M18), and used to predict AD up to M48 (i.e., within 2.5 years of last data collection time). In addition, baseline demographic features were integrated into the feature set. The raw medication and comorbidity modalities were spare, and each data type was carefully preprocessed. For example, patient medications data were prepared based on a specific pipeline of three main steps, namely (1) medical names collected and cleaned, (2) data encoding based on ATC ontology, and (3) semantic aggregation of data to the most appropriate level of granularity. To study the predictive power of these new features, we compared the performance of five ML algorithms: SVM, RF, KNN, DT and LR. These techniques were optimized to perform a 4-class classification task of CN, sMCI, pMCI, and AD. The performance of the models was evaluated using the 10-fold cross-validation technique. Based on the B-CS-AM-D fused dataset, RF achieved state-of-the-art testing

performance (i.e., an accuracy of 90.51%, precision of 90.69%, recall of 90.51%, and F1-score of 90.41%). The models were also evaluated for 3-class and 2-class classification tasks, generally, RF was found to be the most stable and accurate model. Our results show that the fusion of medication and comorbidity features mostly improved the performance of all models. This uncovers the significant role of these cost-effective features in the AD prediction domain. The proposed model is not only efficient and easy to deploy in real medical environments but also medically intuitive and was able to achieve state-of-the-art performance.

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.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2016R1D1A1A03934816).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.future.2020.10.005.

References

[1] Y. Liu, et al., Diffusion tensor imaging and tract-based spatial statistics in Alzheimer's disease and mild cognitive impairment, Neurobiol. Aging 32 (9) (2011) 1558–1571.
[2] World Health OriginatiSon.
[3] J. Neugroschl, S. Wang, Alzheimer's disease: diagnosis and treatment across the spectrum of disease severity, Mount Sinai J. Med.: J. Trans. Personal. Med. 78 (4) (2011) 596–612.
[4] A. Ishiwata, et al., Preclinical evidence of Alzheimer changes in progressive mild cognitive impairment: a qualitative and quantitative SPECT study, Acta Neurol. Scand. 114 (2) (2006) 91–96.
[5] Q. Li, et al., Classification of Alzheimer's disease, mild cognitive impairment, and cognitively unimpaired individuals using multi-feature kernel discriminant dictionary learning, Front. Comput. Neurosci. 11 (2018) 117.
[6] S. Qiu, et al., 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.: Diagn. Assess. Dis. Monit. 10 (2018) 737–749.
[7] C.S. Musaeus, M.S. Nielsen, P. Høgh, Microstates as disease and progression markers in patients with mild cognitive impairment, Front. Neurosci. 13 (2019).
[8] D. Zhang, D. Shen, and A.s.D.N. Initiative, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, NeuroImage 59 (2) (2012) 895–907.
[9] D. Zhang, D. Shen, and A.s.D.N. Initiative, Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers, PLoS One 7 (3) (2012).
[10] C.Y. Wee, et al., Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns, Hum. Brain Mapp. 34 (12) (2013) 3411–3425.
[11] Y. Fan, et al., Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline, NeuroImage 39 (4) (2008) 1731–1743.
[12] 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.
[13] X. Jiang, L. Chang, Y.-D. Zhang, Classification of Alzheimer's Disease via Eight-Layer convolutional Neural Network with batch normalization and Dropout Techniques, J. Med. Imaging Health Inform. 10 (5) (2020) 1040–1048.
[14] Y. Zhang, et al., Multivariate approach for Alzheimer's disease detection using stationary wavelet entropy and predator–prey particle swarm optimization, J. Alzheimer's Dis. 65 (3) (2018) 855–869.
[15] B.A. Ardekani, K. Figarsky, J.J. Sidtis, Sexual dimorphism in the human corpus callosum: an MRI study using the OASIS brain database, Cerebral Cortex 23 (10) (2013) 2514–2520.
[16] K. Shaji, et al., Clinical practice guidelines for management of dementia, Indian J. Psychiatry 60 (Suppl 3) (2018) S312.
[17] B. Cheng, et al., Domain transfer learning for MCI conversion prediction, IEEE Trans. Biomed. Eng. 62 (7) (2015) 1805–1817.
[18] D. Zhang, et al., Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 55 (3) (2011) 856–867.
[19] S. Huang, et al., Identifying Alzheimer's disease-related brain regions from multi-modality neuroimaging data using sparse composite linear discrimination analysis, in: Advances in Neural Information Processing Systems, 2011.
[20] C. Hinrichs, et al., MKL for robust multi-modality AD classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2009.
[21] L. Xu, et al., Multi-modality sparse representation-based classification for Alzheimer's disease and mild cognitive impairment, Comput. Methods Programs Biomed. 122 (2) (2015) 182–190.
[22] K.R. Gray, et al., Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, NeuroImage 65 (2013) 167–175.
[23] P.A. Donnelly-Kehoe, et al., Looking for Alzheimer's Disease morphometric signatures using machine learning techniques, J. Neurosci. Methods 302 (2018) 24–34.
[24] A. Chincarini, et al., Integrating longitudinal information in hippocampal volume measurements for the early detection of Alzheimer's disease, NeuroImage 125 (2016) 834–847.
[25] S. Tabarestani, et al., A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study, NeuroImage 206 (2020) 116317.
[26] L. Huang, et al., Longitudinal clinical score prediction in Alzheimer's disease with soft-split sparse regression based random forest, Neurobiol. Aging 46 (2016) 180–191.
[27] A. Ashfaq, et al., Readmission prediction using deep learning on electronic health records, J. Biomed. Inform. 97 (2019) 103256.
[28] database, T.A.s.D.N.I.A..
[29] A. Niculescu, et al., Blood biomarkers for memory: toward early detection of risk for Alzheimer disease, pharmacogenomics, and repurposed drugs, Mol. Psychiatry (2019) 1–22.
[30] N. Geifman, et al., Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions, Alzheimer's Res. Ther. 10 (1) (2018) 4.
[31] World Health Organisation Collaborating Centre for Drug Statistics Methodology.
[32] S.K. Tayebati, et al., Identification of World Health Organisation ship's medicine chest contents by Anatomical Therapeutic Chemical (ATC) classification codes, Int. Marit. Health 68 (1) (2017) 39–45.
[33] N.V. Chawla, et al., SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.
[34] A.L. Chau, X. Li, W. Yu, Support vector machine classification for large datasets using decision tree and Fisher linear discriminant, Future Gener. Comput. Syst. 36 (2014) 57–65.
[35] C. Apté, S. Weiss, Data mining with decision trees and decision rules, Future Gener. Comput. Syst. 13 (2–3) (1997) 197–210.
[36] M. Shahbaz, et al., Classification of Alzheimer's Disease using Machine Learning Techniques, 2019.
[37] A. 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.
[38] R.E. Wright, Logistic regression, 1995.
[39] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[40] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017.
[41] A. Moscoso, et al., Prediction of Alzheimer's disease dementia with MRI beyond the short-term: Implications for the design of predictive models, NeuroImage: Clin. 23 (2019) 101837.
[42] X. Hong, et al., Predicting Alzheimer's Disease using LSTM, IEEE Access 7 (2019) 80893–80901.
[43] R. Filipovych, C. Davatzikos, and A.s.D.N. Initiative, Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI), NeuroImage 55 (3) (2011) 1109–1119.
[44] E. Westman, J.-S. Muehlboeck, A. Simmons, Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion, NeuroImage 62 (1) (2012) 229–238.
[45] S. Tangaro, et al., A fuzzy-based system reveals Alzheimer's disease onset in subjects with Mild cognitive Impairment, Phys. Med. 38 (2017) 36–44.
[46] P. Moore, et al., Random forest prediction of Alzheimer's disease using pairwise selection from time series data, PLoS One 14 (2) (2019).
[47] D. Yao, et al., An ensemble learning system for a 4-way classification of Alzheimer's disease and mild cognitive impairment, J. Neurosci. Methods 302 (2018) 75–81.

[48] N. Amoroso, et al., Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge, J. Neurosci. Methods 302 (2018) 3–9.
[49] 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.
[50] M. Liu, et al., Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis, IEEE Trans. Biomed. Eng. 66 (5) (2018) 1195–1206.
[51] L. Sørensen, M. Nielsen, and A.s.D.N. Initiative, Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination, J. Neurosci. Methods 302 (2018) 66–74.
[52] S.I. Dimitriadis, et al., 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) database, J. Neurosci. Methods 302 (2018) 14–23.
[53] 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.
[54] 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.

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 as an assistant professor. Currently he is Post-Doctoral Fellow at Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain. 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.

Hager Saleh obtained a master's degree in Computer Science and her bachelor's degree in Information systems from the Faculty of Computer and Information, University Assuit, Egypt. Her research interests are centered on Big data Analytics, Data Mining, Sentiment Analysis, Natural Language Processing, Machine Learning, and Streaming Data.

Radhya Sahal is an Adjunct Lecturer and Postdoctoral Research Fellow in Confirm Centre for Smart Manufacturing, Data Science Institute, National University of Ireland, Galway. Radhya has publications in the area of Big Data and Cloud Database. Her research interests include Big Data, Stream Processing, Healthcare, Internet of Things (IoT), Smart Manufacturing and Smart Cities. She holds her Ph.D. and M.Sc in computer science in 2013 and 2018 respectively from Faculty of Computers and Information, Cairo University, Egypt.

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 at Sungkyunkwan University, South Korea. His research interests include biomedical applications, information security, network security, Internet security, and machine learning and its application to medical, security, and privacy problems.

S. M. Riazul Islam (M'10) received the B.S. and M.S. degrees in Applied Physics and Electronics from University of Dhaka, Bangladesh in 2003, and 2005, respectively and the Ph.D. degree in Information and Communication Engineering from Inha University, South Korea in 2012. He has been working at Sejong University, south Korea as an Assistant Professor at the Department of Computer Science and Engineering since March 2017. From 2014 to 2017, he worked at Inha University, South Korea as a Postdoctoral Fellow at the Wireless Communications Research Center. Dr. Islam was with the University of Dhaka, Bangladesh as an Assistant Professor and Lecturer at the Dept. of Electrical and Electronic Engineering for the period September 2005 to March 2014. In 2014, he worked at the Samsung R&D Institute Bangladesh (SRBD) as a Chief Engineer at the Dept. of Solution Lab for six months. His research interests include wireless communications, 5G & IoT, wireless health, bioinformatics, and machine learning.

Farman Ali is an Assistant Professor in the Department of Software at Sejong University, South Korea. He received his B.S. degree in computer science from the University of Peshawar, Pakistan, in 2011, M.S. degree in computer science from Gyeongsang National University, South Korea, in 2015, and a Ph.D. degree in information and communication engineering from Inha University, South Korea, in 2018, where he worked as a Post-Doctoral Fellow at the UWB Wireless Communications Research Center from September 2018 to August 2019. 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 registered over 4 patents and published more than 50 research articles in peer-reviewed international journals and conferences. He has been awarded with Outstanding Research Award (Excellence of Journal Publications-2017), and the President Choice of the Best Researcher Award during graduate program at Inha University.

Eslam Amer is an associate professor of computer science. Currently, he is working as a postdoctoral research fellow at faculty of electrical engineering and computer science – technical university of Ostrava – Czech Republic. Eslam is working on malware analysis using natural language processing. His main research interests are natural language processing, information retrieval.