Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Development of a three tiered cognitive hybrid machine learning algorithm for effective diagnosis of Alzheimer's disease

Afreen Khan, Swaleha Zubair *

Department of Computer Science, Aligarh Muslim University, Aligarh, India

ARTICLE INFO

Article history:
Received 17 September 2021
Revised 20 June 2022
Accepted 19 July 2022
Available online xxxx

Keywords:
Alzheimer's disease
Cognitive
Feature selection
Hybrid
Machine learning
Mild cognitive impairment

ABSTRACT

Alzheimer's disease (AD) is one of the most frequent neurodegenerative disorders in the elderly subjects. Since early detection can prevent or delay cognitive decline in the older subjects, it is desirable to develop effectual protocols for the diagnosis of the disease. Most of the existing diagnostic tools fail to improvise timely disease prognosis in susceptible patients. Keeping this fact into consideration, we developed a cognitive-based 3-tiered machine learning (ML) algorithm employing baseline characteristics to predict AD or mild cognitive impairment (MCI) to construct psychometric test results. Earlier machine learning based AD diagnosis methods used a binary or multinomial classification technique. We relied on the development of a sophisticated hybrid cognitive ML algorithm that provides an accurate and precise prediction of the disease. We built an ML model using cognitive and demographic data. The prediction method consisted of a three-step process. Alzheimer's Disease Neuroimaging Initiative (ADNI) database was used to develop a novel prediction algorithm. Considering the fact that nineteen ML and deep learning classifiers could not adequately classify ADNI data, we created a 2-layer model stacking procedure. Model stacking outperformed six ML classifier combinations, including Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Trees, Random Forest, and eXtreme Gradient Boosting. The performance of the as-proposed model was evaluated employing seven performance assessment measures and four classification error indicators. Each model was evaluated in three separate strategical assessment modules. In the first experiment, XGB, Random Forest, and SVM achieved 89.63% accuracy, while Random Forest achieved 93.90% accuracy in the second experiment. Experiment 2 improved the classification and performance of overall prediction. In the third experiment, hybrid modeling, the accuracy increased significantly, with experiment 1 giving 90.24% accuracy and experiment 2 yielding 95.12% accuracy. The as-proposed model successfully predicted early AD and MCI in an effective manner. We were able to reduce nineteen classifiers into four classifiers (from experiment-1) and six classifiers (from experiment-2) and subsequently into one *meta*-learner ($19 \rightarrow 4 \rightarrow 1$ and $19 \rightarrow 6 \rightarrow 1$), with high predictive power. Finally, we performed a thorough comparative analysis of different ADNI datasets to validate our findings.

1. Introduction

Alzheimer's disease (AD) is one of the most serious neurological conditions that include cognitive dysfunction in the elderly patients. It was identified by Alois Alzheimer in 1906 as disease condition with criteria of progressive memory loss, disorientation, and pathological indicators (senile plaques and neurofibrillary tangles). This is a global disease affecting the entire world population. According to a general estimate, a new AD case is identified every-three seconds throughout the world. This in turn suggests that around 50 million people are affected globally in a period of around (Harvey et al., 2003; Prince et al., 2016). These estimates are likely to quadruple every twenty years, thereby likely to inflict

* Corresponding author at: Department of Computer Science, Aligarh Muslim University, Adjacent Computer Centre, Anoopshahr Road, Aligarh (202001), U.P., India.

E-mail addresses: afreen.khan2k13@gmail.com (A. Khan), swalehazubair@yahoo.com (S. Zubair).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.07.016
1319-1578/

100 million individuals by 2040. In the latter stages of AD, individuals lose their independence as well as cognitive abilities inevitably. Because early symptoms are akin to healthy aging in many aspects, diagnosis of Alzheimer's can be delayed if cognitive changes are mistaken for aging (Avidan et al., 2009).

Future research endeavours related to advancement in the treatment of patients with Alzheimer's disease include: exploitation of functional brain imaging techniques for early diagnosis and evaluation of treatment efficacy; development of new classes of medications that help in treatment of AD by modulating various facets of neurotransmitter systems (cholinergic, glutamatergic, etc.), or development of drugs that can be used for the treatment of the cognitive deficit and behavioural disturbances; and also in the development of prophylactic methods (amyloid p-peptide immunizations and inhibitors of β and γ-secretase). It is however worth noting that, in addition to age, many of the risk factors associated with diabetes, stroke, cardiovascular disease and metabolic syndrome, play a role in the AD development and progression. The aetiology of hypertension and hyperlipidemia, along with sleep apnea, and systemic dysregulation are considered to be some of the potent risk factors.

The existing approaches to manage AD generally rely on medication and care after the inception of illness. In absence of any effective cure for preventing this important disease, it is imperative that early detection may slow the course of AD thereby delaying the onset of full blown disease.

The early prognosis of AD could be made plausible by the identification of disease-associated reliable markers. In this regard, various neuropsychological, biochemical, and genetic-based markers had been successfully exploited in monitoring dementia progression. Sheehan (2012) implemented a range of assessment scales to determine the severity of dementia. Several small dementia screening tests suited for primary and secondary healthcare have been described in this paper (Sheehan, 2012). These include the Mini Mental State Examination, Abbreviated Mental Test Score, Clock-drawing test, Six-item Cognitive Impairment Test, General Practitioner assessment of Cognition, Mini-Cog, Test Your Memory, Montreal Cognitive Assessment, Addenbrookes Cognitive Assessment and Memory Impairment Screening etc. (Sheehan, 2012). He suggested that the Clinical Dementia Rating, the Global Deterioration Scale, and the Clinicians Global Impression of Change are important in measuring overall dementia severity (Sheehan, 2012). A number of studies have reported the Clinical Dementia Rating (CDR) and the Mini-Mental State Examination (MMSE) as strong predictors of AD-related dementia progression (Nakata et al., 2009; Daly et al., 2000). Wessels et al. (2018) conducted a comparative analysis of the ADAS-Cog (Alzheimer's Disease Assessment Scale – Cognitive Subscale) and the CDR-SB (Clinical Dementia Rating – Sum of Boxes) to unravel treatment group differences in AD (Wessels et al., 2018). They reported that the ADAS cognitive subscale was more frequently used to detect differences in AD treatment groups than the CDR-SB scale (Wessels et al., 2018). Considering the relevance of various AD related clinical parameters in the disease associated parameters, it would be appropriate to seek a cost-effective and easy diagnostic marker that is easily accessible in routine clinical settings.

In the healthcare sector, the implementation of Machine Learning (ML) can provide an effective method of dealing with bulk of information for the accurate diagnosis of the disease. ML, a science of pattern-learning, has the unique ability to deal with bulky datasets leading to the development of précised predictive models (Khan and Zubair, 2018). ML allows an automatic selection of high-value predictors from a pool of possible inputs (Nori et al., 2019). The application of Magnetic Resonance Imaging (MRI) along with the complex ML algorithms has been widely used to distinguish the healthy brain from that of the mildly demented brain (Amoroso, et al., 2017). Battineni et al. (2020) reviewed 435 articles published between 2015 and 2019 that deals with the development of ML based tools to diagnose chronic diseases (Battineni et al., 2020). They finally selected 22 studies to present a comparative analysis (Battineni et al., 2020). The authors further reported dementia as one of the chronic diseases with case-control as a study type, MRI as input feature and Support Vector Machine (SVM) classifier for ML modelling (Battineni et al., 2020). In another study, Battineni et al. (2019) built a ML model for dementia prediction using SVM classifier (Battineni et al., 2019). They performed ML modelling of the longitudinal pool of 150 MRI patients to provide a prediction accuracy of 65.75% (Battineni et al., 2019). Similarly, our group had also developed an improved multi-modal ML pipeline for the prognosis of AD (Khan and Zubair, 2020). The ML program, built on the Random Forest ML classifier to analyse OASIS longitudinal MRI data provided an accuracy of 87.0%, (Khan and Zubair, 2020).

In the past few years, there have been many ML based extensive research programs to predict dementia and AD and its exploitation in early diagnosis of the disease. However, some of these earlier reports relied on the traditional ML classifiers that do not require hyperparameter tweaking or an ensembling technique. This resulted in a model with decreased accuracy and performance. Additionally, no gold-standard algorithm exists for predicting progression in individuals at risk of AD, moreover, the clinical translation appears to be missing. Besides, predictors employed in some of these models may operate as a significant barrier to clinical adoption due to their high cost and/or invasive nature (e.g., lumbar puncture or fludeoxyglucose positron emission tomography scans etc.). Moreover, a suitable sequential data pre-processing pipeline, capable of efficiently addressing issues such as missing data, outliers, and imbalanced classification, is considered as the backbone of every successful ML model (Lim et al., 2007; Balsis et al., 2011). In light of these considerations, developing a system capable of accurately and efficiently predicting Alzheimer's disease and similar disorders, with improved accuracy and performance from clinical, neuropsychological and other existing data poses a substantial challenge.

Cognition is a collection of mental processes that have an impact on practically every aspect of one's life. Cognitive abilities include the ability to reason and to learn new things consistently. Cognitive impairment, on the other hand, is a word used to describe a problem with cognition or reasoning. The severity of the disease varies from moderate to severe, depending on the patient's age and health. Various reasons can lead to cognitive impairment, including age, genetics, and environmental variables. In addition to drug-related adverse effects, blood vessel concerns, depression, and dementia are among the list of potential difficulties (Cognitive Testing, xxxx). Cognitive tests (also referred to as psychometric or neuropsychological assessments) are used to evaluate cognitive impairments in individuals. In general, clinical examinations along with psychometric tests can be used in conjunction to assess the course of MCI and early Alzheimer's disease (Mathotaarachchi et al., 2017).

Considering these challenges, we envisaged the development of a cognitive-based hybrid machine learning model that can be built exclusively on psychometric test scores to predict AD, MCI or cognitively normal (CN) subjects. We proposed an approach for multinomial classification as the most effective and vital form of the classifier in the diagnosis of AD. Multiclass or multinomial classification is a ML problem that involves classifying instances into three or more classes. We developed a model based on three experiments. In the first module, we demonstrated multinomial classification based on feature selection after developing a positive correlation. In the second module, the cognitive features were chosen using a sequential feature selector, which reduces the

y-dimensional feature space to the z-dimensional feature space where y < z. It is a type of greedy search algorithm that was created as a sub-optimal solution. In the final experiment (third module), we built a hybrid cognitive model for each of the two modules (based on experiment 1 and experiment 2) separately. The data used in this study were obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI) database. In the beginning, we employed nineteen ML and deep learning classifiers. As they were unable to correctly categorize ADNI data, therefore, we developed a two-layer model stacking approach. It was devised to test all possible combinations of the nineteen classifiers. Model stacking outperformed on four and six combinations of ML classifiers, including Logistic Regression, Naive Bayes, Support Vector Machine, Decision Trees, Random Forest, and extreme Gradient Boosting. This was the foundation for our experiments 1, 2, and 3.

The ADNI datasets had already been studied by several researchers. In the year 2017, Mathotaarachchi et al. proposed a ML-based predictive model to detect dementia development within 24 months (Mathotaarachchi et al., 2017). They obtained an accuracy of 84.0% and a 0.91 under-receiver operating characteristic curve, respectively (Mathotaarachchi et al., 2017). Grassi et al. (2018) identified MCI individuals at risk of AD conversion using a subset of the ADNI dataset (Grassi et al., 2018). They developed an algorithm using ML techniques with a balanced accuracy of 78.8% (Grassi et al., 2018). Using cascaded multiview canonical correlation, Singanamalli et al. (2017) proposed a classification by integrating a subset of diagnostic modalities (Singanamalli et al., 2017). Their accuracy ranged from 63.0 to 93.0% (Singanamalli et al., 2017). On the ADAS-Cog cognitive test, a maximum accuracy of 93.0% was achieved using a subset of the ADNI dataset with just 149 patients (Singanamalli et al., 2017).

Keeping into consideration the features of the above described approaches, we proposed a three-tiered hybrid strategy for distinguishing CN, MCI, and AD subjects. This approach is based on a combination of neuropsychological tests. In our proposed model, we employed cognitive, clinical, and demographic data to predict the 3-class response variable (CN/MCI/AD). We introduced a comprehensive pipeline procedure for data analysis, transformation, fusion, aggregation and processing for prediction. We established three ways to assess the validity of diagnostic classifications using patient categorization, cohort size imbalances, and cognitive data. It was discovered that a hybrid cognitive model incorporating selected psychometric variables increased the predictive accuracy of AD, MCI, and CN. To further evaluate the effectiveness of our methodology, seven performance evaluation measures and four classification error metrics were used in conjunction with each other. The as-developed procedure yielded efficient and comprehensive diagnostic approaches. To improve clinical practice, our proposed model streamlines the interpretation of test findings by developing a set of criteria to classify the individual and discover three response variables at an early stage.

2. Methods

2.1. Study design

To achieve meaningful and robust diagnostic predictions while dealing with big, multifarious, incomplete, multi-source, and diverse data, the ML approach, proposed in the present study, relied on the proficient data handling, management, processing, aggregation, and synchronization. Further, we tried to improvise on missing pattern detection, data pre-processing, imputation, data transformation and integration (Khan and Zubair, 2020). Following that, we inculcated procedures needed in the automated extraction of structured data from unstructured data. This requires a high throughput and versatile interface to both model-based and model-free techniques that can be applied to multivariate data that has been harmonized and aggregated. This in turn necessitates a high throughput and versatile interface that can be applied to heterogeneous data for the subsequent predictive analytics and diagnostic prognosis.

For the ML classification and modeling, we employed the Python computing environment of Anaconda, where each of the steps were executed, integrated and authenticated. The pipeline environment was established for the complete process. This workflow was employed to assure the success of the whole procedure enable internal verification and provide external reproducible results. Fig. 1 illustrates the flowchart of our end-to-end procedure. In the following subsections, we have presented the introduced approach and various steps involved in this study, in detail.

2.2. Data source

The Alzheimer's Disease Neuroimaging Initiative, ADNI (adni.loni.usc.edu), is a database, launched in 2004 as a public–private partnership, headed by Michael W. Weiner as a principal investigator. It is a longitudinal-based multicentre study that aims in developing biochemical, clinical, genetic, imaging biomarkers for the early detection, progression and tracking of AD and MCI.

We selected ADNIMERGE participant data, which consists of ADNI-1, ADNI-2, ADNI-3 and ADNI-GO series of the database. These were procured at a different phase of the study, each belonging to a distinct time period. In each of the databases, new patients were enlisted while prevailing patients from previous phases remained to be examined. The ADNIMERGE includes 2175 subjects, aged between 54 and 92 years. The data for these participant groups consists of 14,036 input values for 113 features. The input values were acquired for about 8 years (2004–2021), initially with a baseline (when the patient first arrived), and then with a fixed gap of every 6-month follow-up visit, for 8 years, making a total of 14,036 input values. Broadly, this is a dataset that combines significant predictors from all 4 phases, accumulated using several data sources within the ADNI repository.

2.3. Participants

In the present study, we extracted the ADNI-1 data, which consisted of 818 subjects and a total of 5013 input values for 113 features. Next, we built a ML cognitive model, tested and validated it on the basis of as-generated data set, and later compared it with ADNI-2, ADNI-3 and ADNI-GO datasets to ascertain the efficiency of as-developed model results. The 818 participants were populated by demographic information, cerebral spinal fluid, cognitive/neuropsychological/psychometric, diffusion tensor imaging, electroencephalography, genetic, magnetic resonance imaging, and positron emission tomography biomarkers.

Confirmed diagnoses from screening to the baseline visit were included, while more than 20% of missing patient data were excluded. The baseline diagnosis was categorized into five groups, Cognitively Normal (CN), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI), Significant Memory Complaints (SMC), and Alzheimer's Disease (AD), based on their follow-up visit's diagnosis. We grouped them into three categories CN, MCI and AD. The CN group included CN and SMC subjects, the MCI group consisted of EMCI and LMCI subjects while the AD group consisted of the AD subjects. As the main aim of this study was to predict the future decline based on the cognitive assessment, we grouped to stay consistent between diagnoses at distinct instants of time. Out of the 818 participants that were included in the study, 229 subjects were diagnosed as CN, 396 as MCI and 193 were diagnosed with AD. The demographic information of the

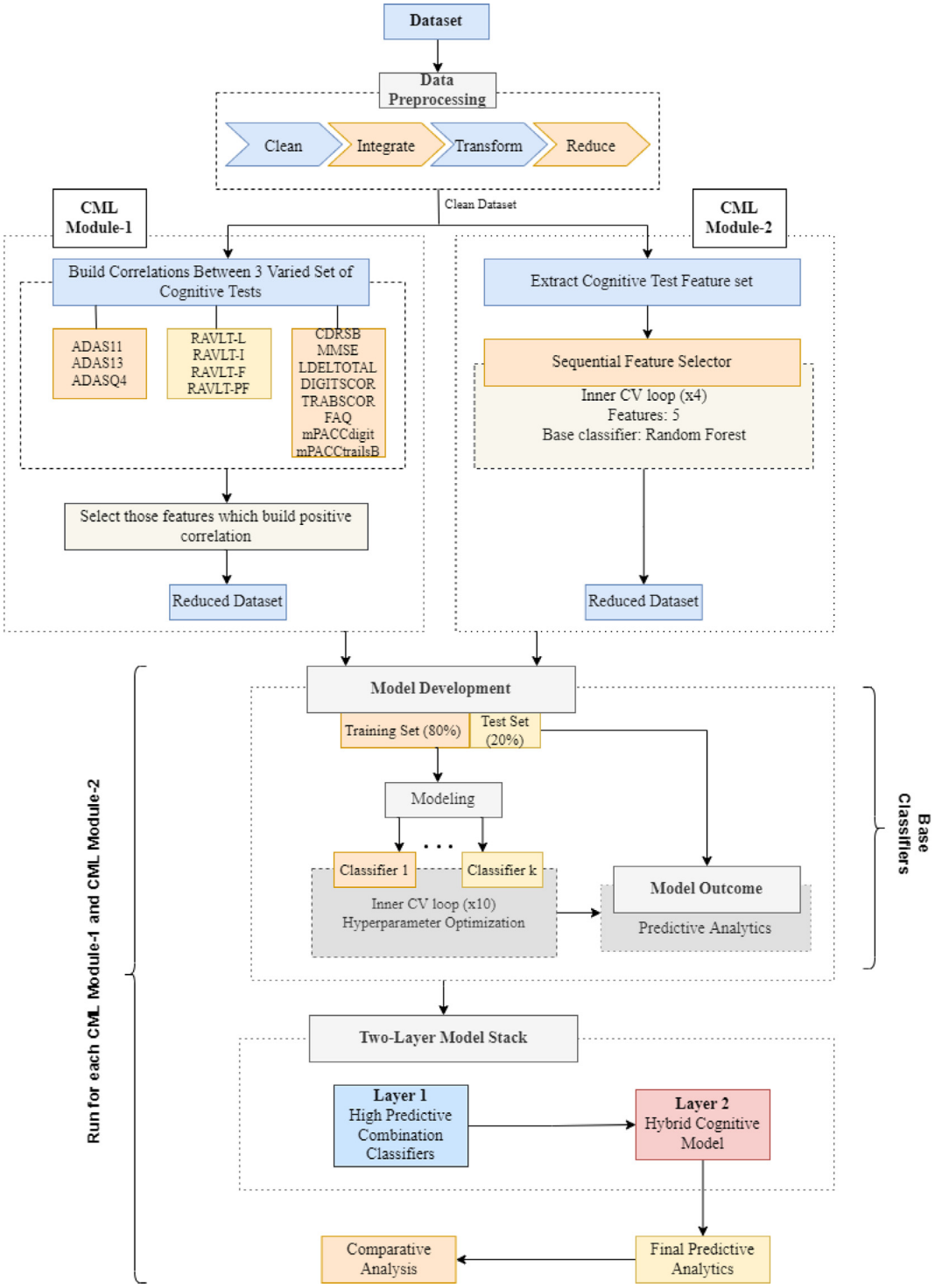

Fig. 1. Cognitive Model Framework. [*CML: Cognitive Machine Learning].

studied participants in accordance with their diagnosis at baseline is presented in Table 1.

2.4. Variable extraction

Taking into consideration the fact that we aimed at using those predictors that are either consistently evaluated or efficiently incorporated into clinical settings, and that are not considered as intrusive by patients, we decided to use only those variables in the ADNIMERGE dataset pertaining to diagnostic subtypes, demographic variables, as well as scores on clinical tests and neuropsychological tests. As in the present study, certain variables were not accessible for all enrolled respondents, it was decided ahead of time to exclude variables with more than 20% missing data. A detailed description of the cognitive tests (clinical and neuropsychological) is presented in Table 2. Table 3 portrays the variables used in the present study. Table 4 gives a comprehensive overview of the cognitive assessment variables present in the dataset,

Table 1
Demographic Details.

Characteristic	All Subjects	CN	MCI	AD
N	818	229	396	193
Gender				
Male	476 (58.20%)	119 (51.96%)	255 (64.39%)	102 (52.84%)
Female	342 (41.80%)	110 (48.03%)	141 (35.61%)	91 (47.15%)
Age				
Range	54.4 – 90.9	59.9 – 89.6	54.4 – 89.3	55.1 – 90.9
Mean (S.D.)	75.18 (6.84)	75.84 (5.02)	74.43 (7.40)	75.28 (7.45)
Years of Education				
Range	4–20	6 – 20	4 – 20	4 – 20
Mean (S.D.)	15.54 (3.05)	16.07 (2.86)	15.63 (3.04)	14.71 (3.13)
Ethnicity				
Non-Hispanic/Latin	793	226	380	187
Hispanic/Latin	19	2	13	4
Unknown	6	1	3	2
Race				
White	761	210	370	181
Black	39	16	15	8
Asian	14	3	9	2
More than 1	3	0	1	2
Indian/Alaskan	1	0	1	0
Marital Status				
Married	630	156	317	157
Widowed	108	40	48	20
Divorced	51	17	25	9
Never Married	28	15	6	7
Unknown	1	1	0	0

Table 2
Cognitive Measures Description.

Variables	Description
ADAS	**Alzheimer's Disease Assessment Scale** A comprehensive examination to assess cognitive and non-cognitive symptoms of AD.
ADAS11	**Alzheimer's Disease Assessment Scale (11 items)** This assessment comprises 11 questions. The number might be anything between 0 and 70. A score of 0 indicates no impairment, while a score of 70 indicates significant impairment.
ADAS13	**Alzheimer's Disease Assessment Scale (13 items)** It includes 13 questions. The value ranges from 0 to 85. A score of 0 indicates no impairment, while a score of 85 indicates significant impairment.
ADASQ4	It is task 4 of ADAS11. It is the word recognition cognitive subscale.
CDRSB	**Clinical Dementia Rating – Sum of Boxes** It measures dementia progression, specifically in individuals with mild to moderate cognitive impairment. A semi-structured interview with the patient and other interviewees (family members) is used to get the rating. The value range is 0 to 18.
DIGITSCOR	**Digit Span Test Score** It is a task that is used to assess the number storage capacity. Participants are given a numerical sequence and are asked to repeat it back to the examiner in either forward span or reverse span.
FAQ	**Functional Assessment Questionnaire** It assesses a patient's ability to carry out everyday activities independently. The scale runs from 0 to 30. A score of 0 indicates normal, whereas a score of 30 indicates that the person is extremely reliant.
LDETOTAL	**Delayed Total Recall** It is a neuropsychological test that evaluates a person's ability to recall information after a prescribed amount of time. It assesses an individual's capability to recall information after a certain time.
MMSE	**Mini Mental State Examination** It is a questionnaire-based assessment, which is used to determine cognitive decline. It has a range of 0 to 30. Normal scores range from 25 to 30, mild scores range from 21 to 24, moderate scores range from 10 to 20, and severe scores range from 0 to 10.
mPACC	mPACC tests assess cognitive abilities, timed executive function and episodic memory.
mPACCdigit	**Modified Preclinical Alzheimer Cognitive Composite with Digit** It is a mPACC test with Digit substitution.
mPACCtrailsB	**Modified Preclinical Alzheimer Cognitive Composite with Trails B** It is a mPACC test with Trails B substitution.
RAVLT	**Rey Auditory Verbal Learning Test** RAVLT is a neuropsychological test that is frequently used to measure auditory-verbal skills such as attentiveness, memory, and learning capacity. The RAVLT is a five-trial procedure (Trials 1–5) that involves presenting a list of 15 words. The subject is asked to recall the terms from the first list again after 30-minutes of interpolated testing. This is called delayed recall. These scores are further used to generate various summary scores.
RAVLT-L	**Rey Auditory Verbal Learning Test – Learning** It is calculated by subtracting Trial 1 score from Trial 5 score.
RAVLT-I	**Rey Auditory Verbal Learning Test – Immediate** It is determined by aggregating the results of the first five trials (Trials 1 to 5).
RAVLT-F	**Rey Auditory Verbal Learning Test – Forgetting** It is calculated by subtracting Delayed Recall score from Trial 5 score.
RAVLT-PF	**Rey Auditory Verbal Learning Test - Percent Forgetting** It is calculated by dividing RAVLT-F score by Trial 5 score.
TRABSCOR	**Trail Making Test Part B Time** It is a diagnostic test that assesses cognitive functioning i.e. the ability to think, reason, and retain information.

Table 3
Dataset Variables.

Demographic	Age, Gender, Education levels, Marital Status
Diagnostic Subtype	CN, MCI, AD
Clinical and Neuropsychological Test	ADAS11, ADAS13, ADASQ4, CDR-SB, RAVLT-L, RAVLT-I, RAVLT-F, RAVLT-PF, MMSE, LDETOTAL, DIGITSCOR, TRABSCOR, FAQ, mPACCdigit, mPACCtrailsB

including their mean and standard deviation (S.D.), as well as the proportion of missing values, which is an important factor to consider for this dataset.

2.5. Data Pre-processing

The acquired dataset was processed using a 4-step strategy i.e., clean, integrate, transform and reduce.

1. **Clean:** In this study, the data was noisy, incomplete and inconsistent. In general, the inaccurate/dirty data causes hindrance in the mining procedure. Since most mining methods include certain techniques, they often deal with missing or noisy data, which is not necessarily resilient. Thus, we run the data through several data cleansing procedures as part of an essential data pre-processing step. This retrospective study required imputing the inconsistent neuroimaging data. Table 4 shows the proportion of missing values for the extracted features at baseline. This missing value problem was fixed using imputation through k nearest neighbours, with k = 5 neighbours. Because the imputed values do not add biases, the outcome of multivariate imputation allows for the analysis of the entire dataset that has a similar joint distribution of the source (original) data (Khan and Zubair, 2020).
2. **Integrate:** Data integration involves combining data from various sources into a cohesive data repository. Multiple databases, data cubes, and flat files are examples of these sources. Redundancy is yet another key consideration. If an attribute is derived from another table, it is sometimes redundant. Redundancies in the ensuing dataset might also be caused by inconsistencies in variables. As the ADNI data warehouse consists of numerous datasets, certain data values that we discovered inconsistent in the ADNIMERGE dataset were extracted from a pool of several other ADNI datasets. The extracted data values were then integrated accordingly, which had the same values for the corresponding subject ID.
3. **Transform:** In this study, data transformation included the following strategies: normalization and smoothing. In normalization, the values were scaled to fall within a defined range. The smoothing was performed to get rid of the noise in the ADNI data.
4. **Reduce:** Data reduction methodologies are effective in analyzing the reduced dataset without affecting the integrity of the original dataset (Khan and Zubair, 2020). Dimension reduction data reduction strategy was employed in this study. It allowed for the detection and removal of variables or dimensions that were irrelevant, weakly related, or duplicated.

2.6. CML Module-1

The objective of this experiment was to examine the classification performance achieved by combining three varied sets of cognitive tests and building a correlation between them. Those psychometric tests that showed a positive correlation were selected. Additionally, this experiment seeks to determine whether this methodology is tailored to optimize the classification and improve performance as compared to the second approach, as described in CML Module-2.

The psychometric tests as described in Table 2 were grouped into three classes. The first class consisted of ADAS tests, the second set included RAVLT tests and the third group comprised remaining cognitive assessments. After correlation analysis, variables were selected that formed the positive correlations. While remaining tests that showed a negative or poor correlation were dropped from the dataset. Thus, RAVLT-L, RAVLT-I, ADAS11, ADAS13, ADASQ4, MMSE, DIGITSCOR, LDELTOTAL, mPACCdigit and mPACCtrailsB cognitive variables were selected for a new reduced dataset.

2.7. CML Module-2

The objective of this experiment was to examine the classification performance based on the cognitive test variables, built after the sequential feature selection was performed. This feature selection algorithm follows the pattern of the greedy search algorithm, which reduces m-dimensional feature space to the *n*-dimensional feature space where m < n. In this process, 4-fold cross-validation and Random Forest as a base classifier were employed. The selection method was standardized to remove irrelevant features and to maximize the model accuracy. Thus, MMSE, CDRSB, RAVLT-I, DIGITSCOR and LDETOTAL were the variables selected and included in the reduced dataset.

Also, this experiment seeks to determine whether this methodology is tailored to optimize the classification and improve perfor-

Table 4
Descriptive Statistics.

S.No.	Variables	CN Mean (S.D.)	MCI Mean (S.D.)	AD Mean (S.D.)	% Missing at Baseline
1.	ADAS11	6.20 (6.20)	11.4 (4.42)	18.60 (6.28)	0.12
2.	ADAS13	9.50 (4.19)	18.62 (6.27)	28.87 (7.62)	0.97
3.	ADASQ4	2.85 (1.72)	6.18 (2.26)	8.56 (1.56)	0.0
4.	CDRSB	0.03 (0.12)	1.60 (0.88)	4.29 (1.64)	0.0
5.	DIGITSCOR	45.75 (10.20)	36.85 (11.17)	26.93 (12.81)	0.61
6.	FAQ	0.14 (0.60)	3.82 (4.46)	12.99 (6.84)	0.36
7.	LDETOTAL	12.97 (3.57)	3.81 (2.27)	1.27 (1.90)	0.0
8.	MMSE	29.11 (0.98)	27.03 (1.78)	23.34 (2.06)	0.0
9.	mPACCdigit	−0.12 (2.47)	−7.47 (3.29)	−13.98 (3.01)	0.0
10.	mPACCtrailsB	−0.33 (2.44)	−7.60 (3.39)	−14.24 (3.09)	0.0
11.	RAVLT-L	5.85 (2.28)	3.30 (2.35)	1.81 (1.79)	0.48
12.	RAVLT-I	43.33 (9.09)	30.76 (9.04)	23.16 (7.70)	0.48
13.	RAVLT-F	3.58 (2.73)	4.67 (2.26)	4.54 (1.91)	0.48
14.	RAVLT-PF	34.18 (27.64)	67.86 (31.41)	88.70 (21.92)	0.97
15.	TRABSCOR	89.21 (44.26)	130.74 (73.69)	197.95 (87.09)	1.71

mance as compared to the first approach, as described in CML Module-1.

2.8. Model development

2.8.1. Classification model

Machine learning and deep learning was used and a classifier was developed that may be used to detect probable underlying instances of AD. A hybrid classification model for Alzheimer's disease was built based on variables that were selected during the CML module-1 and CML module-2. The model development step was run for each of the modules individually. The dataset was divided into training and test set with each encompassing 80% and 20% of the data, respectively. The optimized classifier trained on the complete training set was applied to the independent test set after 10 iterations of 5-fold repeated stratified cross-validation on the training set.

2.8.2. Machine learning and deep learning algorithms

The ADNI dataset was used to train a ML algorithm to differentiate between individuals with AD, MCI and cognitively healthy background. In this study, nineteen machine learning and deep learning based classifiers (described below) were employed to study the impact of the as-built model. Various All analyses were performed in a Python environment, using the implementation of the ML techniques available in the Scikit-Learn library.

A. Ensemble-Based:
   1. AdaBoost (Adaptive Boosting): This ML classifier employs ensembling approach. Ensemble classifiers incorporate a number of ML classifiers. When AdaBoost is used on weak ML classifiers, it yields a strong classifier which is more accurate. As an ensemble classifier, it begins by fitting a ML classifier to the training dataset and then fits multiple replicas of the ML classifier to the same training dataset. The difference is that the weights allocated to misclassified occurrences are changed in such a way that subsequent classifiers put a strong emphasis on challenging cases (Cao et al., 2013). This classifier is most commonly used to enhance the performance of any ML algorithm.
   2. Extra Trees (Extremely Randomized Trees): This learning algorithm is a massively randomised tree classifier that is utilised in the ensemble approaches. It is structured differently from a decision tree classifier. Additionally, they are far faster than Random Forest classifiers. Dissociated Decision Trees are combined in a forest-like structure to produce an outcome for classification purposes. It makes use of a meta-estimator that fits many randomised Decision Trees, namely extra-tress, to various dataset attributes (Geurts et al., 2006). It prevents overfitting and optimises the classifier's accuracy by utilising the averaging method.
   3. Random Forest: It is a meta-estimator ML classifier that uses an ensemble technique. To be accurate, it consists of multiple independent Decision Trees that work in conjunction. It fits a variety of Decision Tree classifiers to subsamples of the original dataset (Denisko and Hoffman, 2018). Every single tree in the Random Forest provides the class prediction. The final resultant model prediction is the class that contains the majority of votes. Additionally, it employs an averaging strategy to improve accuracy and to control over-fitting. This ensemble group then outperforms any of the constituent-specific classifiers. In recent years, it has been utilized more in the identification and categorization of dementia-related issues (Dauwan et al., 2016; Gray et al., 2011).
   4. eXtreme Gradient Boosting (XGB): This classifier utilizes Gradient Boosting in conjunction with a Decision Trees classifier to improve the speed and, as a result, the overall performance of the system (Khan and Zubair, 2018).

B. Tree-Based:
   5. Decision Trees: A Decision Tree is a tree-like structure made up of a root node, an internal node, a leaf node, and a branch. The root node is represented by the top node, the internal node by the features (attributes), the leaf node by the result, and the branch by the decision rule. Additionally, it learns the pattern and splits the data according to the feature values. Decision Tree is a non-parametric supervised machine learning classifier. The primary goal of the decision tree is to construct a model capable of predicting the target feature by learning a set of derived decision rules from the data features (Amancio and Comin, 2014).

C. Generalized Linear Models:
   6. Logistic Regression: It is a statistical technique for predicting the probability of class membership given a set of attribute values (Hosmer et al., 2013). It is a probabilistic model that is used when the dependent variable has a binomial distribution. As a result, the relationship between the dependent and independent variables is defined in terms of a function that can be used to forecast future occurrences. We employed regularisation techniques to prevent the problem of overfitting. Regularization is performed to reduce the cost function to match the parameters to the training data. Lasso and Ridge Regression are two simple strategies for reducing model complexity and avoiding overfitting that can occur with simple Logistic Regression.
      a) Lasso Regression (L1): This is a type of regression model that makes use of L1 regularization. All less significant features are entirely discarded in this method of regularisation when evaluating output. Thus, Lasso Regression can assist users not only in reducing overfitting but also in feature selection. L1 can generate sparse models (models with a small number of coefficients); some coefficients can be zeroed out and omitted.
      b) Ridge Regression (L2): This is a type of regression model that makes use of L2 regularization. This modifies the cost function by adding a penalty equal to the square of the coefficients' magnitude. Thus, ridge regression reduces the size of the coefficients and contributes to the reduction of model complexity and multi-collinearity. L2 models do not produce sparse models and all coefficients are compressed by the same factor (none are discarded).

   7. Passive Aggressive: This is a collection of online learning algorithms that may be used for both machine learning classification and regression. Here, in particular, various algorithms for binary and multinomial classification as well as for regression, sequence prediction, and uniclass prediction are examined (Crammer and Dekel, 2006). This comprehensive analysis enables the identification of the algorithms' worst-case loss constraints. This online classifier is used in conjunction with the partial-fit approach, which trains the model in batches. A HashingVectorizer is used to ensure that the feature space remains constant over time. Each data sample is projected into the uniform feature space by this vectorizer.

8. Stochastic Gradient Descent (SGD): It is a machine learning classifier that is particularly effective for discriminative-based learning of linear models such as Logistic Regression and Support Vector Machines. It incorporates both regularized linear classifiers and SGD learning. Typically, the model it fits is controlled by the loss parameter. The gradient of the loss function is estimated by taking one sample at a time and simultaneously updating the model (Robbins and Monro, 1951). It does, however, require a large number of hyperparameters. For instance, several iterations and parameters for regularisation. In addition, it is hypersensitive to feature scaling, which is one of the primary limitations of the SGD classifier.

D. Naïve Bayes:

9. Naïve Bayes: It is a Bayesian machine learning classification technique based on Bayes' theorem (Rish, 2001). It works by assuming that the effect of a particular feature remains independent of the effect of other feature sets within a given class. This is referred to as class-conditional independence. The following two categories of Naïve Bayes classifiers are described in terms of the distributional assumptions they consider:

    a) Bernoulli Naïve Bayes: This is a special case of the Naive Bayes classifier that is optimized for multivariate modeling. It operates on boolean (binary) features. It implements the Naïve Bayes classifier using Bernoulli distributions with several variables. It considers each feature as binary-valued regardless of the number of features in the training set of data.
    b) Gaussian Naïve Bayes (GNB): This is a variant of the Naive Bayes classifier that is employed when the features are continuous. It is assumed that all feature sets adhere to a Gaussian distribution, i.e. the normal distribution.

E. Neighbor-Based:

10. K Nearest Neighbors (KNN): This method follows an instance-based learning technique. Rather than developing a broad internal model, it captures instances of train data. The KNN classifier considers just those observations that are in close vicinity to the occurrences being predicted (Zhang, 2016). It performs learning centred on the K Nearest Neighbours, wherein K is highly data-dependent and signifies a user-specified numeric value. Additionally, it is a non-parametric machine learning classifier. The term 'non-parametric' refers to the absence of any assumptions about the underlying distribution of the data. It operates well with a limited feature set in comparison to a vast feature set.

F. Support Vector Machine:

11. Support Vector Machine (SVM): These classifiers are a collection of supervised machine learning approaches that may be utilised for classification as well as regression. They execute effectively on large datasets. The support vectors, i.e. decision function, are constructed using a portion of the train data, hence increasing their memory efficiency (Kotsiantis, 2007). It exhibits a wide range of behaviour since the decision function is implemented using a variety of kernel functions. Based on the kernel parameter, the two major types of classifiers are defined for SVM classifiers:
    a) Linear SVM: This ML classifier is similar to the Support Vector Classifier, with the exception that the kernel parameter is set to 'linear'. It is written in liblinear rather than libsvm (Pedregosa, et al., 2011). As a result, it has a greater degree of freedom in selecting an appropriate penalty and loss function parameter among the available penalty and loss function values.
    b) Radial Basis Function (RBF) SVM: The kernel parameter is set to 'rbf' in this type of SVM. RBF adds the normal curves surrounding the data points, allowing the decision boundary to be set by a form of topological criteria, such as curves with a sum greater than 0.5.

G. Neural Network:

12. Perceptron: This is a generalised computational model that is utilised to employ linearly separable functions. It is a ML classifier that is based on the same fundamental notion as the SGD classifier. This is a predictive algorithm that determines the exact class to which a given input belongs (denoted by a vector of numbers). In general, it aggregates the input, i.e. weighted sum, and returns a value of 1 if the weighted total exceeds a threshold value; otherwise, it returns a value 0 (Chawla et al., 2002).

H. Deep Learning:

13. Convolutional Neural Network (CNN): It is a type of deep learning neural network that was developed specifically for the purpose of processing structured arrays of data. It is an algorithm that can take in an image as input, give learnable weights and biases to various aspects/objects in the image, and distinguish between them. A CNN's strength stems from a special type of layer called a convolutional layer. CNN is composed of numerous convolutional layers stacked on top of one another, each capable of identifying more complex structures.
14. Gated Recurrent Unit (GRU): In recurrent neural networks (RNNs), the GRU serves as a gating mechanism. The GRU functions similar to a long short-term memory (LSTM) with a forget gate, but with fewer parameters because it lacks an output gate. It seeks to exploit connections between nodes to accomplish machine learning tasks related to memory and clustering. GRUs aid in the adjustment of neural network input weights in order to address the vanishing gradient problem, which is a prevalent challenge with RNNs.
15. Long Short Term Memory Network (LSTM): The Long Short Term Memory Network (LSTM) is a type of recurrent neural network that is particularly well-suited for dealing with long dependencies, such as those found in sequence prediction problems. LSTMs have feedback connections, which means they can analyse full sequences of data in addition to single data points such as images. A typical LSTM unit is made up of four components: a cell, an input gate, an output gate, and a forget gate. The cell retains data across arbitrary time intervals, and the three gates control the inflow and outflow of information.
16. Deep Neural Network (DNN): It is based on neural network architecture. It is composed of three node layers: input, hidden, and output. It is constructed using a series of fully-connected layers. Each subsequent layer is composed of a collection of nonlinear functions representing the weighted sum of all outputs (fully connected) from the preceding one. It performs classification on the input data via backpropagation. Additionally, the log-loss function is optimized using SGD, or lbfgs, as the solver, which is a type of quasi-Newton technique (Pedregosa, et al., 2011).

2.8.3. Cross-Validation (CV) and data augmentation procedure

The goal was to create an algorithm that can produce the optimum generalized performance instead of only for the instances employed while training. Thus, for each hyperparameter configuration, CV provides an approximation of such generalized performance. In this step, the training data was split into 10 folds of instances that were held-out of the training procedure, with the remaining cases being trained repeatedly. The algorithm was then applied to the held-out samples after they had been trained. A 10-fold repeated stratified CV training and testing approach was implemented.

An imbalanced classification problem is a severe problem in cases where the data distribution is usually biased across the target variable (Kohavi, 1995). It poses a challenge while building a ML model. If not treated well, the imbalance classification leads to the development of a ML model that ignores and result in a poor performing model with lower accuracy. In this study, we resolved the imbalanced classification problem using Synthetic Minority Oversampling Technique (SMOTE). This is a data augmentation technique that specifically deals with minority class (Lavrač et al., 1999). There were three classes of the target variable i.e. normal, MCI, and AD subjects. In particular, there was a disparity found in the MCI and AD classes. Because of this, the built ML model resulted in poor performance and hence, lower accuracy. Hence, we employed SMOTE analysis which oversampled the minority class. Oversampling only balances the class distribution; it does not add any extra information to the ML model. Before testing each classifier, it was tuned with the hyperparameters. Hyperparameters primarily help in structuring the ML model (ML tuning is a kind of optimization problem). Post access to a collection of hyperparameters, we tried to discover the correct combination of their values. This helps in examining the performance metrics of the classifiers with maximum accuracy and other related metrics.

2.8.4. Hyperparameter optimization

ML algorithms often contain one or more hyperparameter that enables the algorithm to be tuned differently throughout the training period. Intending to get the best possible performance when applied to instances that are not part of the training data, varying values of these hyperparameters resulted in algorithms with varied prediction performances. Each model was trained with hyperparameter configurations to optimize such hyperparameters for each ML model applied in this study.

2.8.5. 2-layer model stacking

Since very few of the nineteen base classifiers performed well on the dataset at the model development stage, we constructed a two-layer model stack. This was achieved by making model combinations using stacking. And, this resulted in the formation of four and six model combinations with high predictive power. It was then passed to the hybrid model, which was developed using these classifiers chosen from the nineteen available. Stacking machine learning and deep learning models was performed in layers, depending on the models we have trained and the best combination of these models. The layers of model stacks are depicted in Fig. 2.

Figs. 1 and 2 present the entire execution of the suggested algorithm. We arranged model stacks in layers, and each layer served a different purpose. Level 1 is the first layer which consisted of all the sixteen classifiers, where model stacking was performed with layers. Herein, the model combinations were created. This level was designed to run through all of the possible combinations of nineteen classifiers. Level 1 exhibited significant performance accuracy on the combination of four and six classifiers (each for the CML-1 and CML-2), particularly Logistic Regression, Naïve

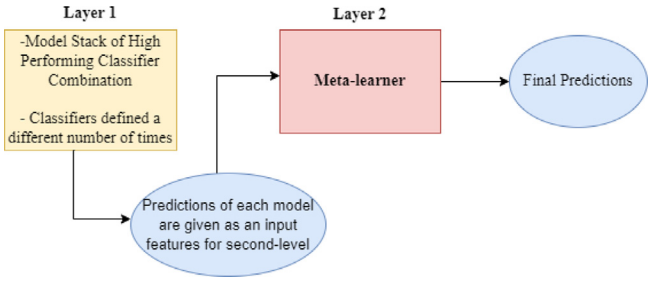

Fig. 2. Model stack layers. (Terms 'level' and 'layer' are interchangeable; their meanings are identical.).

Bayes, Support Vector Machine, Decision Trees, Random Forest, and eXtreme Gradient Boosting. These four and six classifiers were defined a different number of times and layered, resulting in 16 learners for CML Module-1 and 26 learners for CML Module-2. At the second level, we obtained a final dataset, which was then used to create a final model. The final model is known as a meta-learner (hybrid model). Its goal was to incorporate all of the features from each level into the final predictions. This complete model development stage was repeated for each of the two modules separately (CML-1 and CML-2). The hybrid cognitive model for each of the two modules was then developed.

It can be observed from Fig. 2 that each level corresponds to a new layer in our pipeline when model stacking is employed. This is a two-stage pipeline, with level 2 being the final model that provides the final predictions, which serves as the output. Thus, we generate new features through the use of several model stacks, which are then integrated into a new dataset to generate final predictions and, as a result, to develop a meta-learner.

2.9. Model outcome

2.9.1. Predictive analytics

The next step was to make predictions of the likelihood of future outcomes. Both techniques facilitated in classification, prediction, and outcome projection i.e. AD status, through new/test data that had the same clinical data variables that it was trained on.

2.9.2. Validation and inference

To evaluate a learning-based classifier's predictive performance, it is necessary to optimize its accuracy on training data as well as estimate its predictive performance on test data. An optimal classification accuracy depends on many factors including the selection of a classifier, parameter estimation, bias estimates, and precision. Because the intended high precision (low bias) and low variance (high reliability) cannot always be achieved simultaneously, there is frequently a trade-off between absolute prediction accuracy (precision) employing training data with classification reliability (Lavrač et al., 1999; Boustani et al., 2005). Taking this into consideration, a statistical 10-fold CV was used as an alternate approach for validating an estimate and a classification without the requirement for a completely new prospective dataset. Moreover, the entire AD classification model was used to draw inferences with a certain degree of likelihood depending on the findings.

2.10. Performance evaluation

We have a multitude of metrics to help us optimize the model, quantify its performance, compare it, and improve it. In this study, seven performance evaluation measures and four classification errors were used to evaluate the individual model performance.

Different parameters such as accuracy, balanced accuracy, sensitivity/recall, specificity, precision, F-measure, area under the receiver operating characteristic curve, Hamming loss, Jaccard index, Matthews correlation coefficient and logarithmic loss were employed to determine the performance of the multiclass AD classification model.

A. Classification Metrics

It assesses a model's performance and specifies whether the classification is true or untrue, although each of them evaluates it differently. In this study, each parameter was calculated using a 3x3 confusion matrix. When a subject was diagnosed as cognitively normal (CN), it is true positive (TP). It is true negative (TN) when a person was diagnosed with MCI or AD. The predicted value matches the actual value in both cases. But, other times, when the predicted value is falsely predicted, it is classified into false positive (FP) and false negative (FN). The performance metrics for the AD classification model is calculated using a confusion matrix, as illustrated pictorially in Fig. 3.

The descriptions of each of the indicators of the confusion matrix are provided below:

1. **Accuracy (A_{cc}):** It is the percentage of total correct predicted results made from total outcomes.
2. **Balanced Accuracy (BA_{CC}):** When dealing with imbalanced datasets, balanced accuracy in multiclass classification tasks is used. It is simply the mean of sensitivity and specificity.
3. **Sensitivity (S_{en}):** It calculates the percentage of true positives. It determines how many of the actual positive outcomes our model was able to correctly predict i.e. the capability to correctly identify subjects diagnosed with MCI or AD. It is also called recall.
4. **Specificity (S_{pe}):** It measures the proportion of true-negatives i.e. it determines the ability to correctly distinguish subjects diagnosed as normal.
5. **Precision (P_r):** It determines how many of the outcomes that were accurately predicted came out to be positive. Also, it determines whether or not the model is accurate.
6. **F-measure (F_s):** It is regarded as the harmonic mean of the precision and recall of the model. It enables a model to be assessed using a single score that accounts for both accuracy and recall, which is useful for summarising model performance and comparing models. An outcome is a number ranging from 0.0 for the worst F-measure to 1.0 for the best F-measure.
7. **AUC:** It is the area under the Receiver Operating Characteristic (ROC) curve and ranges from 0 to 1. It measures how well a model distinguishes between healthy and non-healthy subjects.

B. Classification Errors:

1. **Hamming loss (H_{Loss}):** It is the percentage of labels that are incorrectly predicted i.e. the proportion of inaccurate labels to total labels. It ranges from 0 to 1, where 0 represents the best value and 1 represents the worst value.
2. **Jaccard index (J_{Index}):** It is the size of the intersection of predicted and true labels to the size of the union of predicted and true labels. It has a value between 0 and 1. The value of 0 represents the worst classification and 1 represents the best classification.
3. **Matthews Correlation Coefficient (MCC):** It accounts for true and false positives and negatives and is often considered as a balanced metric that may be applied even when the classes are of significantly different sizes. It is a correlation coefficient number ranging from −1 to + 1. The perfect prediction coefficient is 1, the average random prediction is 0, and the inverse prediction coefficient is −1.
4. **Logarithmic Loss (L_{Loss}):** It considers the probability behind the models rather than just the classification's final result. The stronger the probability, the better will be the log loss. It quantifies the impurity induced by misclassification. 0 represents a perfect classification with no impurities.

2.11. Hybrid cognitive model (HCM)

In this experiment, the stacked combination at level 1, which resulted in high-performing models were integrated to develop the proposed hybrid/fusion model (level 2). Because of the capacity to construct independent models from training data, we utilized a stacking method in this experiment, as illustrated in Fig. 2. The hybrid model was created for each of the two modules (CML-1 and CML-2). Following that, a comparison study was performed using seven performance assessment criterions and four classification errors. This had the advantage of resulting in high prediction

		PREDICTED CLASS		
		Normal	MCI	AD
ACTUAL CLASS	Normal	1 P_{NN}	2 P_{MN}	3 P_{AN}
	MCI	4 P_{NM}	5 P_{MM}	6 P_{AM}
	AD	7 P_{NA}	8 P_{MA}	9 P_{AA}

$$\text{Acc}\ (\%) = \frac{TP+TN}{TP+TN+FP+FN} \times 100$$

$$\text{BAcc}\ (\%) = \frac{Sensitivity+Specificity}{2} \times 100$$

$$S_{en}\ (\%) = \frac{TP}{TP+FN} \times 100$$

$$S_{pe}\ (\%) = \frac{TN}{TP+FN} \times 100$$

$$P_r\ (\%) = \frac{TP}{TP+FP} \times 100$$

$$F_s\ (\%) = \frac{2*Precision*Recall}{Precision+Recall}$$

True Positive:	**False Positive:**
TP (N) = 1; TP (M) = 5; TP (A) = 9	FP (N) = 4+7; FP (M) = 2+8; FP (A) = 3+6
Aggregate TP = TP (N) + TP (M) + TP (A) / 3	Aggregate FP = FP (N) + FP (M) + FP (A) / 3
True Negative:	**False Negative:**
TN (N) = 5+6+8+9; TN (M) = 1+3+7+9; TN (A) = 1+2+4+5	FN (N) = 2+3; FN (M) = 4+6; FN (A) = 7+8
Aggregate TN = TN (N) + TN (M) + TN (A) / 3	Aggregate FN = FN (N) + FN (M) + FN (A) / 3

Fig. 3. 3x3 Confusion matrix illustrations with performance evaluation metrics calculation formulas.

accuracy. Furthermore, integrating several models can result in noise reduction, reduced bias, and improved predictions. All the above-described approaches also evaluate the impact of the chosen method on the findings. To meet these objectives, we compare the results of all three experiments with that of other ADNI datasets, later in the study.

3. Results

In this section, we study the early diagnosis of AD, according to the results of performed ML modeling and predictive analytics as described in Section 2.

The performance accuracy and other related metrics of the initial ML and deep learning modeling, executed for the nineteen classifiers (level 1), are presented in Table 5 for CML Module-1 and Table 6 for CML Module-2.

From Table 5 and Table 6, it can be comprehended that for both of the modules, the designed algorithm did not perform efficiently on all the classifiers. As already described in Section 2.8.5, since these classifiers alone were not able to operate effectively on the ADNI data, we designed a 2-layer model stacking during the model development stage. As Fig. 2 illustrates, in the level 1, we stacked high performing classifiers. This gave us the different stacking combinations, starting from a combination of 2 through a

Table 5
Performance Measures on Base Classifiers for CML Module-1.

S.No.	Model	A_{cc} (%)	BA_{cc} (%)	AUC	S_{en} (%)	P_r (%)	F_s
Machine Learning Algorithms							
1.	AdaBoost	75.60	79.23	0.922	75.60	79.58	0.753
2.	Extra Trees	88.41	86.19	0.966	88.41	88.83	0.886
3.	Random Forest (RF)	89.63	88.48	0.962	89.63	90.16	0.898
4.	eXtreme Gradient Boosting (XGB)	89.63	88.48	0.970	89.63	90.16	0.898
5.	Decision Tree (DT)	86.00	83.01	0.936	86.00	86.48	0.862
6.	Lasso Logistic Regression (LR - L1)	87.20	84.47	0.966	87.20	87.34	0.871
7.	Ridge Logistic Regression (LR - L2)	89.02	87.33	0.969	89.02	89.49	0.892
8.	Passive Aggressive	77.44	77.42	0.956	77.44	82.97	0.768
9.	Stochastic Gradient Descent	88.41	84.05	0.898	88.41	88.69	0.882
10.	Bernoulli Naïve Bayes (BNB)	65.24	61.51	0.773	65.24	72.26	65.36
11.	Gaussian Naïve Bayes (GNB)	81.70	82.20	0.953	81.70	82.44	0.816
12.	K-Nearest Neighbors	85.36	84.12	0.943	85.37	85.48	0.854
13.	Linear SVM	86.00	85.85	0.965	86.00	86.50	0.861
14.	RBF SVM	89.63	87.19	0.961	89.63	89.71	0.896
15.	Perceptron	81.09	77.86	0.883	81.00	84.04	0.810
Deep Learning Algorithms							
16	Convolutional Neural Network (CNN)	87.19	84.68	0.922	83.54	85.25	0.839
17.	Gated Recurrent Unit (GRU)						
a.	Activation = tanh, Recurrent activation = Sigmoid, Optimizer = Adam	11.50	10.57	0.640	57.34	56.23	0.573
b.	Activation = tanh, Loss = Binary_crossentropy, Optimizer = Adam	36.00	34.24	0.726	67.45	62.34	0.660
18.	Long Short-Term Memory (LSTM)	89.02	85.95	0.952	87.80	88.62	0.879
19.	Deep Neural Network						
a.	Solver = Adam, Alpha = 1e-5, Hidden layer size = (15,1)	83.60	83.60	0.957	86.00	86.67	0.862
b.	Solver = lbfgs, Activation = tanh, Alpha = 1e-5, Hidden layer size = (15,1)	88.41	86.05	0.967	88.41	87.00	0.885

Table 6
Performance Measures on Base Classifiers for CML Module-2.

S.No.	Model	A_{cc} (%)	BA_{cc} (%)	AUC	S_{en} (%)	P_r (%)	F_s
Machine Learning Algorithms							
1.	AdaBoost	90.85	91.87	0.960	90.80	92.48	0.911
2.	Extra Trees	92.60	89.63	0.992	92.68	92.60	0.925
3.	Random Forest (RF)	93.90	96.74	93.91	93.90	93.91	0.939
4.	eXtreme Gradient Boosting (XGB)	93.29	96.36	93.27	93.29	93.27	0.933
5.	Decision Tree (DT)	93.30	91.24	0.988	93.30	93.46	0.933
6.	Lasso Logistic Regression (LR - L1)	88.41	85.34	0.973	88.41	88.41	0.884
7.	Ridge Logistic Regression (LR - L2)	92.68	90.22	0.990	92.68	92.75	0.927
8.	Passive Aggressive	75.60	74.39	0.940	75.6	82.00	0.745
9.	Stochastic Gradient Descent	86.00	86.81	0.927	86.00	87.21	0.861
10.	Bernoulli Naïve Bayes (BNB)	80.48	75.63	0.882	80.48	81.05	0.810
11.	Gaussian Naïve Bayes (GNB)	91.00	88.90	0.981	90.85	91.06	0.909
12.	K-Nearest Neighbors	87.00	82.23	0.961	87.00	88.00	0.864
13.	Linear SVM	90.24	87.00	0.973	90.24	90.53	0.901
14.	RBF SVM	90.00	87.00	0.972	90.24	91.00	0.901
15.	Perceptron	84.14	75.00	0.871	84.14	87.00	0.822
Deep Learning Algorithms							
16.	Convolutional Neural Network (CNN)	90.24	88.94	0.962	92.73	92.29	0.920
17.	Gated Recurrent Unit (GRU)						
a.	Activation = tanh, Recurrent activation = Sigmoid, Optimizer = Adam	7.00	6.45	0.568	45.23	43.12	0.452
b.	Activation = tanh, Loss = Binary_crossentropy, Optimizer = Adam	36.00	35.00	0.713	72.34	74.14	0.712
18.	Long Short-Term Memory (LSTM)	92.68	89.63	0.952	92.68	92.65	0.986
19.	Deep Neural Network						
a.	Solver = Adam, Alpha = 1e-5, Hidden layer size = (15,1)	85.36	81.22	0.962	85.36	86.53	0.853
b.	Solver = lbfgs, Activation = tanh, Alpha = 1e-5, Hidden layer size = (15,1)	91.46	89.00	0.986	91.46	91.45	0.914

combination of 19. Although there were many high performing classifiers, but a suitable combination of few classifiers only incremented the overall accuracy while building the hybrid model. For four of the ML classifiers' combinations for module-1 and six of the ML classifiers' combinations for module-2, the model stacking provided us with greater accuracy. This constituted the basis of our experiments 1, 2, and 3. In the following sub-sections, we present the findings of these four and six ML classifiers. And thereafter, we report the results of the hybrid model in experiment 3, for each of the two modules.

3.1. Experiment 1: AD prediction with CML Module-1

This phase classifies subjects into three categories: normal, MCI and AD, based on features selected after building correlations between three groups of cognitive tests. Table 7 reports the performance outcomes of the four combinations of models.

We can comprehend from Table 7 that the XGB model produced better performance as compared to the other tested models in terms of A_{cc}, BA_{CC}, and AUC. Also, XGB has the highest sensitivity (recall), specificity, precision and F-score, followed by RBF SVM. The least performance for S_{en}, S_{pe}, P_r, and F_s was shown by Decision Trees.

The highest accuracy of 89.63%, balanced accuracy of 88.48% and AUC 0.970 was shown by XGB respectively. As previously stated, the AUC is used in diagnostic evaluations to distinguish true state subjects and determine the optimal cut-off values. Furthermore, a larger AUC predicts AD better in given subjects. As a result, with an AUC of 0.970, the XGB model correctly differentiated between true AD individuals. According to the model performance evaluation, as shown in Table 7, the comparison of data classification accuracy, balanced accuracy for diagnosis of the response variable is depicted in Fig. 4(a)-(b).

In the diagnosis of a disease, each model has a level of inaccuracy. Hamming loss, Jaccard index, Matthews correlation coefficient, and log loss were used to determine such errors. The values of these errors in the diagnosis of Alzheimer's disease are shown in Table 7. It can be seen from Table 7, that the H_{Loss} is lowest for the XGB algorithm while Decision Trees has the highest error in H_{Loss} criterion. Based on these results, we can say that the error shown for H_{Loss} for all the four models is much better, as the values are close to 0. This signifies that the least fraction of targets were misclassified.

Concerning the J_{Index}, the XGB and RBF SVM algorithm has the least error and Decision Trees has the highest degree of error. As J_{Index} compares the predicted and true classes, based on these values from Table 7, it can be comprehended that the value closer to 1, shown by XGB and RBF SVM models has the best classification. The J_{Index} value of 0.754 shown by DT is not that worse, as it is closer to 1. So, we can say that the degree of error for J_{Index} is considerably better for all the four models in the diagnosis of AD. Furthermore, XGB exhibits the lowest error in the MCC criterion as compared to other models. A value of 0.835, which is closer to 1, ascertains the best prediction. All the other three models follow XGB in having less error, as the values are closer to 1, showing a significant prediction rate. Finally, concerning L_{Loss} criterion, Decision Trees has the lowest error as compared to other algorithms. And, following Decision Trees, Lasso Logistic Regression, XGB and RBF SVM had similar levels of error, with the lowest error in dementia diagnosis.

Table 7
Performance results determined from the confusion matrix for experiment-1.

Model				
	LR - L2	RBF SVM	DT	XGB
Classification Metrics				
A_{cc} (%)	89.02	89.63	86.00	89.63
BA_{cc} (%)	87.33	87.19	83.01	88.48
AUC	0.969	0.961	0.938	0.970
S_{en} (%)	89.02	89.63	86.00	89.63
S_{pe} (%)	94.86	84.02	92.61	94.61
P_r (%)	89.49	89.71	86.48	90.16
F_s	0.892	0.896	0.862	0.898
Classification Errors				
H_{Loss}	0.109	0.104	0.140	0.103
J_{Index}	0.802	0.812	0.754	0.812
MCC	0.825	0.834	0.775	0.835
L_{Loss}	0.277	0.313	0.181	0.289

3.2. Experiment 2: AD prediction with CML Module-2

The classification results and performance evaluation for CML module 2 (as mentioned in Section 2) is presented here. This experiment classifies subjects into three categories: normal, MCI and AD, based on sequential feature selection where cognitive test variables were selected. Table 8 shows the model performance outcomes of the six models.

As shown in Table 8, Random Forest generated a high accuracy of 93.90% for the AD diagnosis, whereas SVM gave the lowest accuracy of 90.24%. Also, Random Forest has the highest sensitivity and specificity as well as precision and F-score, followed by Decision Trees and XGB. SVM showed the lowest performance for S_{en}, S_{pe}, P_r, and F_s. Furthermore, Decision Trees, XGB, Logistic Regression, and Naive Bayes all had classification accuracy of 93.30%, 93.29%, 92.68%, and 90.85%, respectively. In terms of balanced accuracy, similar performance was noticed with a higher value of 91.93% for Random Forest while a lowest of 87.0% for SVM. When compared to AD prediction using CML module-1, this approach demonstrated improvements in terms of accuracy, balanced accuracy, and AUC.

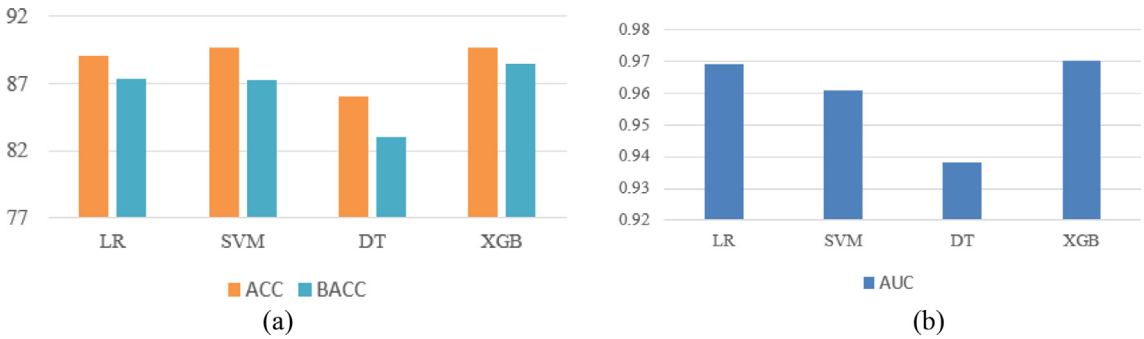

Fig. 4. Performance comparison based on accuracy and AUC for experiment-1.

Table 8
Performance outcomes for experiment-2 as measured from the confusion matrix.

Model	LR - L2	GNB	Linear SVM	DT	RF	XGB
Classification Metrics						
A_{cc} (%)	92.68	91.00	90.24	93.30	93.90	93.29
BA_{cc} (%)	90.22	88.90	87.00	91.24	91.93	90.78
AUC	0.990	0.981	0.973	0.988	0.983	0.993
S_{en} (%)	92.68	90.85	90.24	93.3	93.9	93.29
S_{pe} (%)	95.98	95.24	95.06	96.26	96.74	96.36
P_r (%)	92.75	91.06	90.53	93.46	93.91	93.27
F_s	0.927	0.909	0.901	0.933	0.939	0.933
Classification Errors						
H_{Loss}	0.073	0.091	0.098	0.067	0.061	0.067
J_{Index}	0.863	0.832	0.822	0.874	0.885	0.874
MCC	0.882	0.853	0.844	0.893	0.902	0.892
L_{Loss}	0.158	0.238	0.273	0.141	0.34	0.152

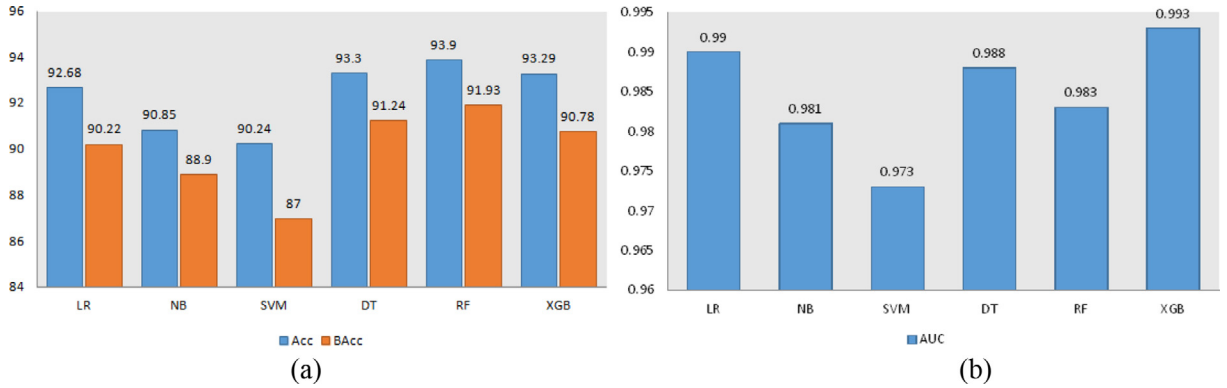

Fig. 5. Performance comparison based on accuracy and AUC for experiment-2.

According to the model performance evaluation, as indicated in Table 8, Fig. 5(a)-(b) illustrates the comparison of data classification accuracy, balancing accuracy, and diagnostic accuracy for the response variable. Fig. 5 shows that XGB has the highest AUC, indicating that it is a better diagnostic predictor, with a 0.993 score, followed by Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, and SVM, which have 0.990, 0.988, 0.983, 0.981, and 0.973, respectively.

Table 8 also presents the errors identified in the AD multiclass classification using four criteria: Hamming loss, Jaccard index, Matthews correlation coefficient, and log loss. It shows that the Random Forest algorithm has the lowest H_{Loss}, whereas SVM has the largest error in the H_{Loss} metric. Based on these findings, we can conclude that the H_{Loss} error for all six models is significantly reduced as compared to experiment 1, with values near 0 and less than 0.1. This means that just a small percentage of targets were misclassified. In terms of J_{Index}, the Random Forest method produces the least error, whereas SVM produces the most. All of the methods have a degree of inaccuracy that is almost identical to that of a random forest. Based on the values in Table 8, it is clear that the value closest to 1 provided by the Random Forest model has the best classification. The J_{Index} value of 0.822 shown by SVM is not worse, as it is near to 1. As a result, we may conclude that the degree of error for J_{Index} is significantly enhanced for all the six models in the diagnosis of AD.

In comparison to other algorithms, the Random Forest has the lowest error of 0.902 in the MCC criteria. In terms of the level of error, random forest, Decision Trees, XGB, Logistic Regression, Naïve Bayes, and SVM are all closer together. A greater MCC metric value indicates that we achieved a much-improved prediction since the values are closer to 1. Finally, in terms of the L_{Loss} criteria, Decision Trees have the lowest error, whereas the Random Forest model has the highest, as can be seen from Table 8. As a result, the degree of impurity induced by misclassification was higher in the Random Forest and lower in the Decision Trees.

The present experiment's results, which outperformed the previous one in terms of prediction accuracy, stimulated the development of new approaches for optimizing prediction accuracy. As a result, we expanded our research to investigate the results of hybrid modeling using CML modules 1 and 2.

3.3. Experiment 3: AD prediction with hybrid cognitive model (HCM)

In this analysis, a hybrid ML model was built, each with CML Module-1 and CML Module-2, as explained in Section 2.12. Since we constructed a prediction model based on the high-performing combination of four and six algorithms in the preceding part, we began to improve our outcomes by ensembling certain models. In this, we establish two levels, the first of which has four models in case of CML-1 and six models in case of CML-2 (layer 1) and the second of which contains a meta-learner (layer 2), as illustrated in Fig. 2. In the following sub-sections, we present the results of the hybrid modeling, trained and tested each with the two modules.

3.3.1. HCM with CML Module-1 (HCM-1)

As shown in Fig. 6, we constructed a first level in which each of the four ML models was defined a different number of times, resulting in a total of 16 learners. It yielded a fresh training set for the second level model comprising of predictions from the first level model. Finally, the second level model contained the

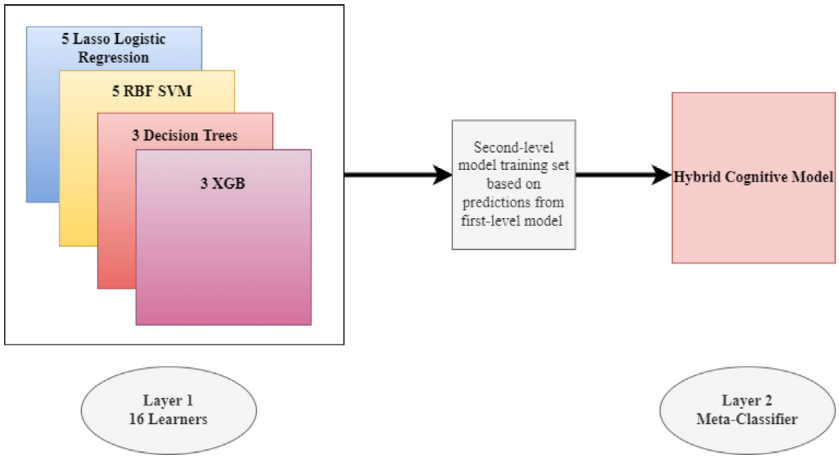

Fig. 6. Learners in hybrid model based on experiment-1.

Table 9
Performance statistics of hybrid model with experiment-1.

Model	S_{en} (%)	S_{pe} (%)	P_r (%)	F_s
Hybrid with CML Module-1	90.24	94.73	90.55	0.903
	A_{cc} (%)	**BA_{cc} (%)**	**AUC**	
	90.24	88.21	0.968	
	H_{Loss}	**J_{Index}**	**MCC**	**L_{Loss}**
	0.098	0.822	0.844	0.281

meta-classifier, from which the final predictions and performance evaluation were determined. This is the final class prediction of the hybrid model.

The performance of the hybrid model with CML module-1 is presented in Table 9. The classification accuracy, balanced accuracy and AUC rates produced were 90.24%, 88.21% and 0.968. Moreover, the sensitivity, specificity, precision and F-score achieved were 90.24%, 94.73%, 90.55% and 0.903. For Hamming loss, Jaccard index, MCC, and log loss criterion, the rates of error in the AD diagnosis were 0.098, 0.822, 0.844, and 0.281, respectively. In terms of the degree of error, we can conclude that we improved H_{Loss} significantly, as the number is now closer to 0, indicating that only a small percentage of targets were misclassified. The error rate for J_{Index} is 0.812, which is close to 1, suggesting that this model did significantly better classification. The MCC metric score is closer to 1, implying that a near-perfect prediction was made once again. Finally, the L_{Loss} value denotes the presence of a little impurity produced by the misclassification.

3.3.2. HCM with CML Module-2 (HCM-2)

In this, we created a first level in which each of the six machine learning models was specified a different number of times, yielding a total of 26 learners (Fig. 7). It resulted in a new training set for the second level model, which included predictions from the first. Eventually, the second level model comprised the fusion model, which is the meta-learner, from which the final predictions and performance evaluation were produced.

Table 10 shows the performance of the hybrid model with CML module-2. Classification accuracy, balanced accuracy, and AUC scores were 95.12%, 94.23%, and 0.990, respectively. Furthermore, the sensitivity, specificity, precision, and F-score attained were 95.12%, 97.50%, 95.20%, and 0.952. The error rate for Hamming loss, Jaccard index, MCC, and log loss criteria were 0.05, 0.907, 0.922, and 0.150. To sum up, we can say that our H_{Loss} value was optimal, meaning that only a small number of targets were incorrectly classified. The error rate for J_{Index} is 0.907, which is close to one, indicating that this model best evaluated the predicted and true classifications of AD, MCI, and cognitively normal. An accurate forecast is indicated by an MCC metric score nearing 1. As a case of impurity, the L_{Loss} value of 0.150, which is closer to 0, indicates a minor misclassification caused by this model.

The parameters for the built hybrid predictive models are given in Table 11.

We may conclude from these findings that the hybrid modeling has allowed us to attain maximum AD subject predictions without bias. We were able to reduce nineteen classifiers into a combination of four and six classifiers and then to a one meta-learner (19 → 4 → 1 and 19 → 6 → 1), that had a high predictive power amongst all. In this study, as per the evaluation of performance disparities in the preceding experiments, the hybrid modeling approach with limited cognitive features has shown to be an effective method in AD-related research.

4. Comparative analysis

The prediction accuracy of the first two tests, with CML module-1 and CML module-2, is shown in Fig. 8. Fig. 9 compares the hybrid model's prediction accuracy for two modules. Based on Fig. 8, it is clear that CML module-2 had a high predictive power on the six classifiers that later formed the basis of hybrid modeling. As mentioned in Section 3 that there were several high performing classifiers out of the nineteen base classifiers.

But a suitable combination of few classifiers only incremented the overall accuracy while building the hybrid model. This is the reason why the comparative analysis of these six classifiers is reported in this section.

CMLM-1 generated 89.63% accuracy for XGB and SVM, whereas CMLM-2 generated a higher accuracy of 93.90% for the Random Forest. Whereas, Fig. 9 shows that the fusion modeling using CMLM-2 was able to predict AD in older individuals with 95.12% accuracy. As a result of the findings, joint modeling with few variables appears to be the most effective method for predicting AD onset. Essentially, this demonstrates that classification performance based on cognitive test variables, which was constructed following the sequential feature selection that served as the foundation for the CMLM-2 experiment, resulted in improved performance on both the six classifiers and the hybrid modeling.

Amongst all, the hybrid modeling with experiment 2 produced the most improved outcomes. This implies that the model built with sequential feature selection, five cognitive variables, cross-validation of four and Random Forest as a base classifier outperformed the one built with correlation analysis. A simulation of

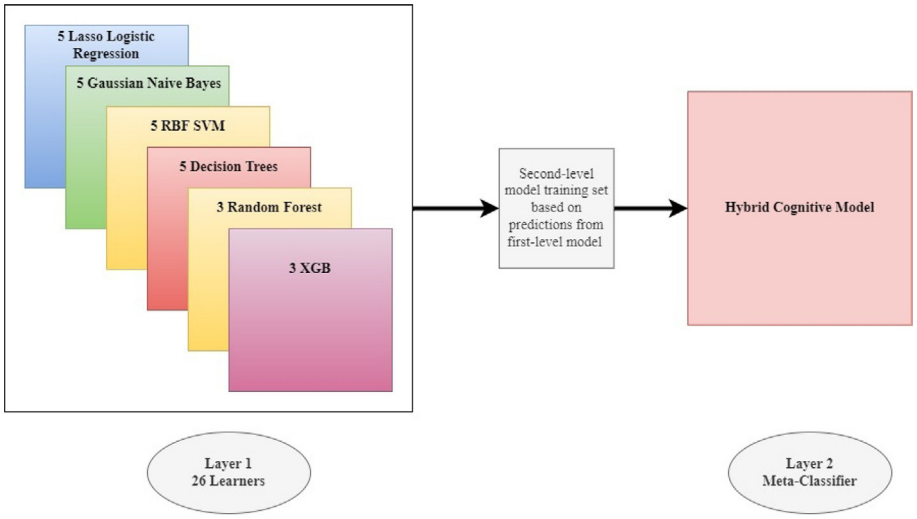

Fig. 7. Learners in hybrid model based on experiment-2.

Table 10
Performance statistics of hybrid model with experiment-2.

Model	S_{en} (%)	S_{pe} (%)	P_r (%)	F_s
Hybrid with CML Module-2	95.12	97.50	95.20	0.952
	A_{cc} (%)	**BA_{cc} (%)**	**AUC**	
	95.12	94.23	0.990	
	Hamming Loss	**Jaccard Index**	**MCC**	**Log Loss**
	0.05	0.907	0.922	0.150

Table 11
Model parameters.

Model	HCM-1		HCM-2	
	Number of learners	16	Number of learners	26
LR	5	penalty = 'l2', C = 1.0, solver = 'sag', multi-class = 'auto', max_iter = 100	5	penalty = 'l2', C = 1.0, solver = 'newton-cg', multi-class = 'auto', max_iter = 100
GNB	–	–	5	None
SVM	5	C = 1.0, kernel = 'rbf', degree = 3, gamma = 'scale'	5	C = 1.0, kernel = 'linear', degree = 3, gamma = 'scale'
DT	3	criterion = 'gini', splitter = 'best', max_depth = 8/6, min_samples_split = 5, min_samples_leaf = 5	5	criterion = 'gini', splitter = 'best', max_depth = 8/5, min_samples_split = 2, min_samples_leaf = 1
RF	–	–	3	n_estimators = 1000, criterion = 'gini', min_samples_split = 2, max_features = 'auto
XGB	3	objective = 'multi:softprob', num_class = 10	3	objective = 'multi:softprob', num_class = 10, learning_rate = 0.1

six enrolled models was run, and the AUC of the model was approximately equal to one as a result of this technique.

To summarize, in experiment 1, XGB and SVM had the greatest accuracy of 89.63%, followed by Logistic Regression, and Decision Trees. At 86.00% accuracy, Decision Trees had the lowest accuracy rate. In experiment 2, Random Forest outperformed other classifiers for MCI and AD patients, scoring 93.90%, followed by Decision Trees, XGB, Logistic Regression, and Naïve Bayes. SVM had the lowest accuracy, at 90.24%. With a 95.12 percent accuracy, the hybrid model with CML module-2 performed best in experiment 3.

We validated our results by executing the as-developed ML model on the remaining ADNI datasets i.e. ADNI-2, and ADNI-3. Following that, a comparison was conducted between ADNI-2, ADNI-3, and the above-built model on ADNI-1 data. On the contrary, other datasets of ADNI, ADNI-GO dataset includes only MCI patients' data. Therefore, we could not employ our algorithm on this dataset. Table 12 summarizes the entire analysis. Table 12 illustrates that the developed model performed better on the other two ADNI datasets as well. The outcomes of the ADNI-2 and ADNI-3 fusion models demonstrated that the model we developed using ADNI-1 data is factual and comprehensible.

5. Comparison with the State-of-the-Art

In absence of effective cure of the Alzheimer's disease, early diagnosis is critical because it may provide an opportunity to the clinicians to opt for the preventive measures. In general, the diagnosis of the disease is carried out in the primary care settings,

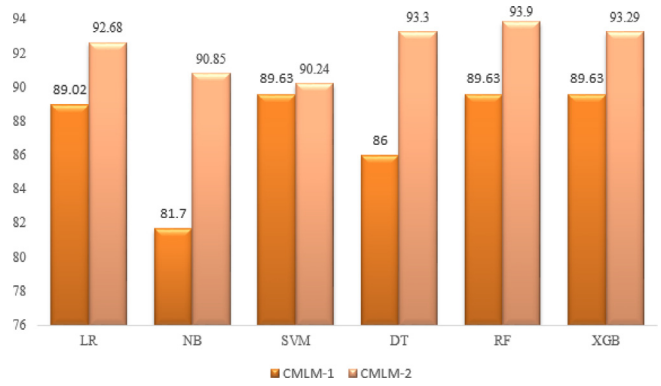

Fig. 8. Comparison of prediction accuracy (%).based on module-1 (CMLM-1) and module-2 (CMLM-2).

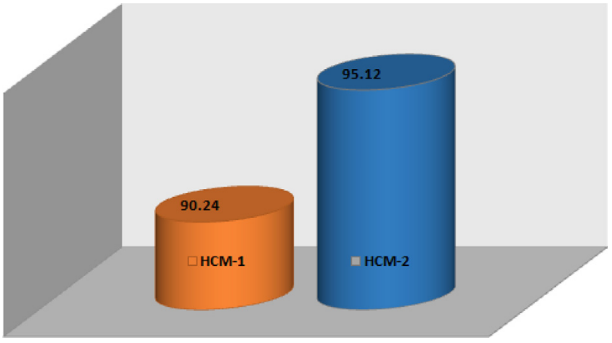

Fig. 9. Prediction accuracy (%) comparison of hybrid model-1 (HCM-1) and hybrid model-2 (HCM-2).

which lack specialist imaging equipment (Beltrán et al., 2020). This may cause overlooking and delayed diagnosis of AD related ailments. Although there are areas where deviations from normal settings may result in interpretable biomarker signals, external disturbances, differences due to many factors that are only marginally associated with AD risk, and also collinearities between biomarkers, typically present challenges in developing explicit biomarkers (Battineni et al., 2020). However, ML models are widely used in real-time clinical practice, as well as in diagnostic and Alzheimer's treatment (Khan et al., 2021). Multiple MRI studies have been incorporated into machine learning models for AD prediction (Khan and Zubair, 2020; Garrard et al., 2014; Liu et al., 2020; Gopi et al., 2020; Bin-Hezam and Ward, 2019). Nonetheless, there is no complete model that can enhance model accuracy concerning cognitive assessments. In light of this, we developed a hybrid model based on neuropsychometric testing to improve the accuracy of AD identification. Table 13 presents the comparison between the proposed approach and similar studies that have been done before.

Table 13 indicates that the proposed model significantly outperformed other models. We must also observe that there is a substantial degree of variability in the datasets, the number of subjects in each study, the classifiers employed, and the modeling approach, thereby, making a direct comparison extremely challenging. Venugopalan et al. (2021), Tanveer et al. (2022), Hazarika et al. (2021), Razzak et al. (2022), and Benge et al. (2009) performed the deep learning modeling using MRI scans, whereas we analyzed AD on non-image data.

6. Significance of the proposed approach

According to our findings, a few biomarkers combined with ML algorithms can provide results almost as excellent as those obtained from more expensive imaging techniques such as MRI and PET, which should still be utilized as a final validation. These less expensive, less invasive biomarkers also happen to consume far less patient and staff time during the acquisition step, allowing for mass screening and participation from sites with modest resources. Because they are less costly and less intrusive, these biomarkers also need considerably less time from patients and personnel during the collection phase, allowing for mass screening and participation from places with minimal resources.

The importance of hybrid ML modeling to forecast Alzheimer's disease in the elderly has been shown in this study. Three separate experiments were carried out, each focused on the construction of correlations and the sequential selection of features. Using conventional diagnostic methods, a diverse set of independent cognitive characteristics were utilized to identify the AD group. Moreover, six predictive models were employed, and the findings show that the prediction accuracy of each model improves over time. A hybrid model each for two modules was developed, where comprehensible results were seen.

Thus, the findings of this work are comparable with those of the above-mentioned researchers (Section 5) who have demonstrated that machine learning and deep learning methods may be utilized

Table 12
Comparative analysis of our proposed approach on the ADNI-1 (A-1) dataset with that of ADNI-2 (A-2) and ADNI-3 (A-3) dataset.

Experiment	Model	Accuracy			AUC			Sensitivity			Specificity		
		A-1	A-2	A-3	A-1	A-2	A-3	A-1	A-2	A-3	A-1	A-2	A-3
CML MODULE-1	LR	87.20	79.62	**81.72**	0.966	0.949	0.926	87.20	79.62	81.72	92.96	88.41	88.09
	GNB	82.70	73.25	78.31	0.953	0.916	0.909	81.70	73.25	78.31	90.77	82.32	86.46
	SVM	87.20	**81.53**	80.72	0.965	0.956	0.920	87.20	81.53	80.72	92.29	89.60	87.03
	DT	86.00	79.61	78.31	0.938	0.901	0.906	86.00	79.62	78.31	92.61	89.24	87.00
	RF	88.41	80.89	80.71	0.963	0.937	0.924	88.41	80.89	80.71	94.12	89.28	88.09
	XGB	**89.63**	81.20	80.72	0.969	0.933	0.918	89.63	82.17	80.72	94.61	89.01	88.09
Hybrid	HCM-1	**90.24**	**83.17**	**82.00**	0.970	0.943	0.927	90.24	83.17	82.00	94.73	89.64	89.63
CML MODULE-2	LR	92.68	**93.63**	**90.36**	0.990	0.922	0.985	92.68	93.63	90.36	95.98	96.52	94.57
	GNB	90.85	91.08	90.36	0.981	0.983	0.983	90.85	91.08	90.36	95.24	95.22	95.36
	SVM	90.24	**93.63**	89.16	0.973	0.992	0.951	90.24	93.63	89.16	95.06	96.37	95.08
	DT	93.30	93.00	88.00	0.988	0.991	0.975	93.30	93.00	88.00	96.26	95.90	94.01
	RF	**93.90**	**93.63**	**90.36**	0.983	0.993	0.988	93.90	93.63	90.36	96.74	96.49	95.10
	XGB	93.29	93.01	91.56	0.993	0.992	0.991	93.29	93.00	91.56	96.36	96.11	95.64
Hybrid	HCM-2	**95.12**	**94.27**	**92.77**	0.995	0.994	0.997	95.12	94.27	92.77	97.50	96.75	96.19

The bold values mean the highest value (accuracy) for the particular experiment performed, for each of the A-1, A-2, and A-3 datasets.

Table 13
Comparative analysis of our proposed approach with other related work on the ADNI dataset.

Author (s)	Highlight	Subjects	Accuracy (%)/AUC
Bin-Hezam and Ward (2019)	Developed how machine learning algorithms can aid in the prediction of dementia based on an individual's profile of modifiable risk factors.	All ADNI cohorts except ADNI3. A total of 1812 subjects were considered in the analysis.	91.53 (dementia vs non-dementia), 77.0 (AD, MCI, and CN) using Logistic Regression.
Cohen (2019)	Aimed to develop a deep learning-based technique for the classification of multi-categorical AD data.	800 subjects. (200 CE, 400 MCI, 200 CN).	87.20 (AD, MCI, and CN) for ANN. 88.28 (AD, MCI, and CN) for 1D-CNN.
Albright (2019)	Leveraged machine learning and neural networks to predict the progression of Alzheimer's disease and also proposed a preprocessing algorithm.	Train data: 1737 subjects. Test data (separate): 110 subjects.	0.866 AUC (AD, MCI, and CN) using Multilayer Perceptron.
Ashraf et al. (2021)	Classified Alzheimer's disease using a variety of CNN-based transfer learning algorithms.	378 subjects. (94 AD, 138 MCI and 146 CN).	99.05 (AD, MCI, and CN) using DenseNet.
Venugopalan et al. (2021)	Conducted an integrated analysis of MRI, genetic, and clinical test data in order to classify patients into AD, MCI, and cognitively normal.	3315 subjects. Clinical: 2004 (707 CE, 699 MCI and 598 CN). Imaging: 502 (266 CE, 104 MCI and 132 CN). Genetic: 809 (226 CE, 338 MCI and 226 CN).	87.00 (AD), 88.00 (CN), 80.00 (MCI) using a combination of Deep Learning and Random Forest.
Tanveer et al. (2022)	Developed a deep learning-based architecture (an ensemble of deep NNs) called as Deep Transfer Ensemble, that was trained using transfer learning for AD classification.	813 3D-MRI scans. (187 CE, 398 MCI, 228 CN).	99.05 and 85.27 on two independent splits for CN vs AD. 98.71 and 83.11 for MCI vs AD, on two independent splits.
Hazarika et al. (2021)	Presented deep learning modeling, where the original DenseNet-121 architecture's convolution layers were replaced with depth-wise convolution layers for AD classification.	210 subjects. 15120 MRI scans. (70 AD, 70 MCI, 70 CN).	90.22 (AD, MCI, and CN) using DenseNet-121 with depth-wise convolution layers.
Razzak et al. (2022)	Developed a multi-resolution PartialNet ensemble optimized for Alzheimer's detection.	350 subjects. 3925 MRI scans. (95 AD, 146 MCI, 95 CN).	98.23 (AD, MCI, and CN).
Proposed Approach	**Developed a 3-tiered cognitive hybrid machine learning algorithm for AD prediction.**	**818 subjects (ADNI-1). (193 CE, 396 MCI, 229 CN).**	**90.24 (AD, MCI, and CN) with HCM-1. 95.12 (AD, MCI, and CN) with HCM-2.**

to diagnose AD. Our approach in the diagnosis of Alzheimer's disease may be comparable to those stated in terms of the employed ML techniques. However, our suggested model, which is based on the approach employed in Alzheimer's support centres' for early diagnosis, cannot only detect AD using data from standard tests from patient populations but also achieve greater accuracy in early diagnosis of AD and MCI.

The efficacy of cognitive scores as a marker of disease severity in Alzheimer's patients has been challenged several times (Benge et al., 2009). As a result of our research, we observed that certain neuropsychological tests were positively associated with other neuropsychological variables. Based on this analysis, we were able to further develop the model. The achieved results of the AD diagnosis broadly represent variability in cognitive test's performance. We did discover, however, that certain of the individual features of the baseline neuropsychological scores can predict AD and MCI progression effectively.

7. Limitations and future scope

Although the developed cognitive ML model performed better than other traditional models in terms of accuracy, the present study has some limitations. Due to the limited number of subjects studied, the final AD, MCI, and CN subject prediction to the total subjects of the ADNI dataset might have been affected.

There is a wide range of improvements that can be made to this dataset and our approach that will be beneficial for future study. By facilitating individualized data-driven medicine, the ability to simulate the stochastic disease progression of particular patients in high resolution might have a profound influence on patient treatment. As a consequence, each patient with a certain disease has distinct risks and thus, probabilistic models cannot generate subject-level predictions with high certainty because of this variability. Therefore, determining variance estimates along with model predictions is also essential when using data-driven techniques for customized medicine and clinical decision support systems.

Artificial intelligence-assisted brain studies may help accelerate current neurological research. Meanwhile, to minimize data constraints, it would be imperative to increase the studied sample size in future research. Simultaneously, hybrid modeling can also be used on younger patients or those with moderate AD, as well as additional biological tests, such as diffusion tensor imaging, cerebrospinal fluid, or other biomarkers to predict accuracy. As surrogate biomarkers are becoming more prevalent in clinical trials, including more varied data into our model development will be our future step.

8. Conclusions

We developed an end-to-end methodology for data analysis, transformation, data fusion, aggregation and processing as well as diagnostic predictions in this work. Based on patient categorization, cohort size imbalances and cognitive data, we developed three techniques to test the validity of diagnostic classifications. Using MRI results, we constructed several ML models to predict AD/MCI/CN in the elderly. It was discovered that a hybrid cognitive model with chosen psychometric features improved the accuracy

of AD and MCI prediction. The first experiment was focused on feature selection by building a positive correlation model. Using a sequential feature selector, a greedy search method was used in the second experiment to choose the robust features. We developed a two-layer stack procedure to determine high-performing stacked combinations of ML classifiers. It was possible to achieve an accuracy of 89.63% using four ML algorithms and a thorough pipeline in the first trial, which used XGB, and 93.90% using the Random Forest in the second using six ML algorithms. Experimentally, we noticed that experiment 2 increased categorization and performance over experiment 1. This number increased considerably when hybrid modeling was used, yielding 90.24% accuracy using experiment 1 and 95.12% accuracy using experiment 2. The prediction models established in this work anticipate the onset of early Alzheimer's disease and MCI. In this work, we conducted a comparative analysis of other ADNI datasets to validate our findings.

To enhance clinical practice, our suggested model simplifies the interpretation of test results by creating a set of criteria to classify the patient and identify AD and MCI at an early stage, utilizing cognitive and demographic data. It is critical to emphasize that every prediction, even those made by ML algorithms, is probabilistic and will always have some degree of error. But, at the same time, the benefit of using algorithmic decision-making tools is that these discrepancies are specified by a definite and empirically examined level of confidence. However, to provide assurance, an algorithm must undergo many testing phases before its use can be securely suggested. We intend to enhance the accuracy of a model that can predict AD more precisely by incorporating several related and unrelated factors in the future.

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.

Acknowledgement

The data used in this study were acquired from the Alzheimer's Disease Neuroimaging Initiative (adni.loni.usc.edu) database.

References

Albright, J., 2019. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. Alzheimers Dement (N Y). 25 (5), 483–491. https://doi.org/10.1016/j.trci.2019.07.001. PMID: 31650004; PMCID: PMC6804703.

Amancio, D.R., Comin, C.H., et al., 2014. A systematic comparison of supervised classifer - supporting information. PLoS ONE 9 (4), 1–13.

Amoroso, N., et al., 2017. Brain structural connectivity atrophy in Alzheimer's disease, arXiv:1709.02369 [physics.med-ph], pp. 1–16.

Ashraf, A., Naz, S., Shirazi, S.H., Razzak, I., Parsad, M., 2021. Deep transfer learning for alzheimer neurological disorder detection. Multimed. Tools Appl. 80 (20), 30117–30142.

Avidan, M., Searleman, A., Storandt, M., Barnett, K., Vannucci, A., Saager, L., Xiong, C., Grant, E., Kaiser, D., Morris, J., Evers, A., 2009. Long-term cognitive decline in older subjects was not attributable to noncardiac surgery or major illness. Anesthesiology 111 (5), 964–970.

Balsis, S., Miller, T.M., Benge, J.F., Doody, R.S., 2011. Dementia staging across three different methods. Dement. Geriatr. Cogn. Disord. 31 (5), 328–333.

Battineni, G., Chintalapudi, N., Amenta, F., 2019. Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). Informatics Med Unlocked 16, 100200.

Battineni, G., Sagaro, G.G., Chinatalapudi, N., Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. J. Pers. Med. 10, 21.

Beltrán JF, Wahba BM, Hose N, Shasha D, Kline RP, For the Alzheimer's Disease Neuroimaging Initiative (2020) Inexpensive, noninvasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's Disease Neuroimaging (ADNI) database. PLoS ONE 15(7): e0235663. https://doi.org/10.1371/journal.pone.0235663

Benge, J.F., Balsis, S., Geraci, L., Massman, P.J., Doody, R.S., 2009. How well do the ADAS-Cog and its subscales measure cognitive dysfunction in Alzheimer's disease? Dementia Geriatr. Cogn. Disorders 28, 63–69.

Bin-Hezam R. and Ward T.E., A Machine Learning Approach towards Detecting Dementia based on its Modifiable Risk Factors, International Journal of Advanced Computer Science and Applications (IJACSA), 10(8), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100820

Boustani, M., Callahan, C.M., Unverzagt, F.W., Austrom, M.G., Perkins, A.J., Fultz, B.A., et al., 2005. Implementing a screening and diagnosis program for dementia in primary care. J. Gen. Intern. Med. 20 (7), 572–577. https://doi.org/10.1111/j.1525-1497.2005.0126.x. PMID: 16050849.

Cao, Y., Miao, Q.G., Liu, J.C., Gao, L., 2013. Advance and prospects of AdaBoost algorithm. Acta Autom. Sin. 39 (6), 745–758.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. 16, 321–357.

Cognitive Testing, Medline Plus. Available online: https://medlineplus.gov/lab-tests/cognitive-testing/.

Cohen, D.S. et al., 2019. Alzheimer's Disease Neuroimaging Initiative. Deep learning-based classification of multi-categorical Alzheimer's disease data. Curr . Neurobiol. 10 (3), 141–147.

Crammer, K., Dekel, O., et al., 2006. Online passive-aggressive algorithms Koby. J. Mach. Learn. Res. 7, 551–585.

Daly, E., Zaitchik, D., Copeland, M., Schmahmann, J., Gunther, J., Albert, M.S., 2000. Predicting conversion to Alzheimer disease using standardized clinical information. Arch. Neurol. 57 (5), 675–680.

Dauwan, M., Zande, J.J., Dellen, E., Sommer, I.E.C., Scheltens, P., Lemstra, A.W., Stam, C.J., 2016. Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease. Alzheimers Dement. (Amst) 4 (1), 99–106.

Denisko, D., Hoffman, M.M., 2018. Classification and interaction in random forests. PNAS 115 (8), 1690–1692.

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., Gorno-Tempini, M.L., 2014. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. Cortex 55, 122–129.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

Gopi, B., Nalini, C., Francesco, A., 2020. Late-Life Alzheimer's Disease (AD) Detection Using Pruned Decision Trees. Int. J. Brain Disord. Treat. 6, 033.

Grassi M, Rouleaux N, Caldirola D, Loewenstein D, Schruers K, Perna G, Dumontier M; Alzheimer's Disease Neuroimaging Initiative. A Novel Ensemble-Based Machine Learning Algorithm to Predict the Conversion From Mild Cognitive Impairment to Alzheimer's Disease Using Socio-Demographic Characteristics, Clinical Information, and Neuropsychological Measures. Front Neurol. 2019 Jul 16;10:756. doi: 10.3389/fneur.2019.00756. PMID: 31379711; PMCID: PMC6646724.

Gray K, Aljabar P, Heckemann R, et al. et al. Random forest-based manifold learning for classification of imaging data in dementia. In: Suzuki K, Wang F, Shen D, eds. Machine Learning in Medical Imaging. MLMI 2011. Lecture Notes in Computer Science. Volume 7009. Berlin: Springer. 2011; 159–166.

Harvey, R.J., Skelton-Robinson, M., Rossor, M.N., 2003. The prevalence and causes of dementia in people under the age of 65 years. J. Neurol. Neurosurg. Psychiatry 74, 1206–1209.

Hazarika, R.A., Kandar, D., Maji, A.K., 2021. An experimental analysis of different Deep Learning based Models for Alzheimer's Disease classification using Brain Magnetic Resonance Images. J. King Saud Univ. – Comput. Inf. Sci. https://doi.org/10.1016/j.jksuci.2021.09.003.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons, Hoboken, NJ.

Khan, A., Zubair, S., 2020. Expansion of Regularized Kmeans Discretization Machine Learning Approach in Prognosis of Dementia Progression. In: 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225397.

Khan, A., Zubair, S., 2018. Machine Learning Tools and Toolkits in the Exploration of Big Data. Int. J. Comput. Sci. Eng. 6 (12), 570–575.

Khan, A., Zubair, S., 2020. An Improved Multi-Modal based Machine Learning Approach for the Prognosis of Alzheimer's Disease. J. King Saud Univ. - Comput Inf. Sci.

Khan, A., Zubair, S., 2020. A Machine Learning-based robust approach to identify Dementia progression employing Dimensionality Reduction in Cross-Sectional MRI data. First International Conference of Smart Systems and Emerging Technologies (SMARTTECH) 2020, 237–242. https://doi.org/10.1109/SMART-TECH49988.2020.00060.

Khan, A., Zubair, S., Khan, S., 2021. Comprehensive Performance Analysis of Neurodegenerative disease Incidence in the Females of 60–96 year Age Group. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 10 (2) https://doi.org/10.14201/ADCAIJ2021102183196.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int. Joint Conference on Artificial Intelligence (IJCAI), 1137–1143.

Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. Informatica 31, 249–268.

Lavrač, N., Flach, P., Zupan, B., 1999. Rule evaluation measures: A unifying view. Springer.

Lim, W.S., Chong, M.S., Sahadevan, S., 2007. Utility of the Clinical Dementia Rating in Asian Populations. Clin. Med. Res. 5 (1), 61–70.

Liu, L., Zhao, S., Chen, H., Wang, A., 2020. A New Machine Learning Method for Identifying Alzheimer's Disease. Simul. Model. Pract. Theory 99, 102023.

Mathotaarachchi, S., Pascoal, T.A., Shin, M., Benedet, A.L., Kang, M.S., Beaudry, T., Fonov, V.S., Gauthier, S., Rosa-Neto, P., 2017. Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiol. Aging 59, 80–90.

Nakata, E., Kasai, M., Kasuya, M., Akanuma, K., Meguro, M., Ishii, H., Yamaguchi, S., Meguro, K., 2009. Combined Memory and Executive Function Tests Can Screen Mild Cognitive Impairment and Converters to Dementia in a Community: The Osaki-Tajiri Project. Neuroepidemiology 33 (2), 103–110.

Nori, V.S., Hane, C.A., Martin, D.C., Kravetz, A.D., Sanghavi, M., 2019. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS ONE, 1–15.

Prince, M.; Comas-Herrera, A.; Knapp, M.; Guerchet, M.; Karagiannidou, M. World Alzheimer Report 2016: Improving Healthcare for People living with Dementia. Coverage, Quality and Costs Now and in the Future; Alzheimer's Disease Int.: London, UK, 2016; pp. 1–140. Available online: https://www.alz.co.uk/research/world-report-2016.

Razzak, I., Naz, S., Ashraf, A., Khalifa, F., Bouadjenek, M.R., Mumtaz, S., 2022. Mutliresolutional ensemble PartialNet for Alzheimer detection using magnetic resonance imaging data. Int. J. Intell. Syst., 1–18 https://doi.org/10.1002/int.22856.

Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence; IBM: New York, NY, USA, 2001.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 22 (3), 400–407.

Sheehan, B., 2012. Assessment scales in dementia. Therap. Adv. Neurol. Disorder 5 (6), 349–358.

Singanamalli, A., Wang, H., Madabhushi, A., et al., 2017. Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via Fusion of Clinical. Imaging and Omic Features. Sci Rep 7, 8137. https://doi.org/10.1038/s41598-017-03925-0.

Tanveer, M., Rashid, A.H., Ganaie, M.A., Reza, M., Razzak, I., Hua, K.-L., 2022. Classification of Alzheimer's Disease using ensemble of deep neural networks trained through transfer learning. IEEE J. Biomed. Health. Inf. 26 (4), 1453–1463. https://doi.org/10.1109/JBHI.2021.3083274.

Venugopalan, J., Tong, L., Hassanzadeh, H.R., et al., 2021. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci. Rep. 11, 3254. https://doi.org/10.1038/s41598-020-74399-w.

Wessels, A.M., Dowsett, S.A., Sims, J.R., 2018. Detecting Treatment Group Differences in Alzheimer's Disease Clinical Trials: A Comparison of Alzheimer's Disease Assessment Scale - Cognitive Subscale (ADAS-Cog) and the Clinical Dementia Rating - Sum of Boxes (CDR-SB). J. Prev. Alzheimer's Dis. - JPAD 5 (1), 15–20.

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Ann. Transl. Med. 4 (11), 1–7.

Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: The Journal of Machine Learning Research 12 (2011), pp. 2825–2830. url: http://dl.acm.org/citation.cfm?id=2078195 (visited on 02/10/2015).