Received: 4 November 2019 | Revised: 16 January 1800 | Accepted: 28 January 2020 | Published online: 2 July 2020

DOI: 10.1002/trc2.12049

Alzheimer's & Dementia
Translational Research & Clinical Interventions

RESEARCH ARTICLE

Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests

Abhay Gupta[1] | **Bratati Kahali**[2]

[1] Undergraduate Program (Physics), Indian Institute of Science, Bengaluru, Karnataka, India

[2] Centre for Brain Research, Indian Institute of Science, Bengaluru, Karnataka, India

Correspondence
Bratati Kahali, Centre for Brain Research, Indian Institute of Science, Bengaluru, Karnataka, India 560012.
E-mail: bratati@iisc.ac.in

Abstract

Introduction: An extensive battery of neuropsychological tests is currently used to classify individuals as healthy (HV), mild cognitively impaired (MCI), and with Alzheimer's disease (AD). We used machine learning models for effective cognitive impairment classification and optimized the number of tests for expeditious and inexpensive implementation.

Methods: Using random forests (RF) and support vector machine, we classified cognitive impairment in multi-class data sets from Rush Religious Orders Study Memory and Aging Project, and National Alzheimer's Coordinating Center. We applied Fisher's linear discrimination and assessed importance of each test iteratively for feature selection.

Results: RF has best accuracy with increased sensitivity, specificity in this first ever multi-class classification of HV, MCI, and AD. Moreover, a subset of six to eight tests shows equivalent classification accuracy as an entire battery of tests.

Discussions: Fully automated feature selection approach reveals six to eight tests comprising episodic, semantic memory, perceptual orientation, and executive functioning can accurately classify the cognitive status, ensuring minimal subject burden.

KEYWORDS
cognitive impairment/dementia, Fisher's score, machine learning, multi-class, neuropsychological tests

1 | BACKGROUND

Dementia, including Alzheimer's disease (AD), is a neurological syndrome which currently affects more than 40 million people globally.[1] These numbers have more than doubled from 1990 to 2016.[2] More and more cases are being reported from low- and middle-income group countries, like India and China, in recent years.[3] AD represents the primary cause of neurodegenerative dementia. AD pathology is defined by cognitive impairment, behavioral disturbance, and functional disabilities, which have a great impact on the quality of daily life of the patient and are a major problem for families and caregivers, too.[4,5] Advancement in prevention of dementia is warranted; however, this is challenging as widespread early detection or screening is necessary. This is achievable partially by extensive assessment at population level via research or community studies. One of the inexpensive ways to do this is neuropsychological assessment, which plays a crucial role in detecting loss of cognitive functions and change in behavioral and functional state compared to normal conditions. Neuropsychological tests can measure different cognitive domains (eg, language, learning, and memory) and subdomains (eg, long-term memory and

recognition memory), and often there is a large battery of tests that needs to be undertaken by the individual.[6] However, the assessment is time-consuming and exhausts the participant, and in this diagnostic process the judgment is also subjective. In the past, there have been studies to demonstrate how decision rules can guide the clinical diagnosis of AD,[7] considering neuropsychological assessments, and pathological features, such as amount of tangles and plaques.

Our aim is to assess the potential of machine learning (ML) approaches in classifying cognitive status. We also aim to identify the best combination of a subset of neuropsychological tests to be used for identifying individuals as cognitively healthy (HV), mild cognitively impaired (MCI) or affected with AD that can be implemented accurately, yet expeditiously in the community. Using ML algorithms on training and testing data sets, we classify individuals as HV, MCI, and AD, while considering demographic data, such as patient's age and education level, for the entire battery as well as the subset of tests. This approach, when implemented successfully, would enable larger community-based studies of aging at reduced costs.

ML models are extensively used to analyze large, complex medical data sets,[8] and integrative analysis of biomedical data.[9] ML algorithms carefully learn the relationships between input variables (eg, test scores) and response variables (eg, clinical diagnosis), and thereby successfully detect disease status of individuals or classify data sets based on their characteristic features. There have been studies on optimizing the combination of neuropsychological tests using ML models, but the studies have been done only for a small data set and involve feature reduction guided by neuropsychologists.[10] In this article, we extensively used and compared different ML approaches to identify individuals as HV, MCI, and AD on large data sets from different study centers, while optimizing for the combination of a subset of significantly decisive tests from the entire neuropsychological battery.

RESEARCH IN CONTEXT

1. **Systematic review**: Machine learning is popularly used in neuroscience, especially neuroimaging, for detecting signs of early neurodegeneration in longitudinal studies. Earlier, a study used feature reduction, subsequently validated by neuropsychologists, to identify four cognitive tests out of twelve as frequent best predictors in binary classifications for three-class Clinical Dementia Rating scores.
2. **Interpretation**: Our findings demonstrate that 8 tests on a battery of 24 accurately (95%) predict the cognitive status of an individual. These tests span episodic, semantic memory, perceptual orientation, and executive functioning domains. This is the first report of a fully automated multi-class prediction with unanimous results from multiple neuropsychological data sets.
3. **Future directions**: This manuscript depicts a framework for the precise classification of healthy, mild cognitive impairment, and Alzheimer's disease individuals ensuring minimal subject burden. Future studies could potentially implement these automated tests in research or community screening. These tests, in conjunction with genetic and/or biological markers, could augment the classification towards absolute accuracy.

2 | METHODS

2.1 | Data sets used

We used neuropsychological tests and their scores from the Rush Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP)[11,12] and the National Alzheimer's Coordinating Center (NACC)[13] for the purpose of our study. Both of these studies have a well-defined dictionary to link the data sheet to the type of tests administered along with the age at visit and years of education information of the individuals at their baseline, which we included in our working models. Moreover, we normalized the tests scores according to the maximum score possible in a test as available from their dictionaries.

2.1.1 | Rush Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) study data

We used the neuropsychological test scores and demographic data (age at visit and years of education) of the individuals from their baseline visit. These individuals were diagnosed with six different types of outcome, which are the following: (1) NCI: no cognitive impairment; (2) MCI: mild cognitive impairment, no other condition contributing to cognitive impairment (CI); (3) MCI+: mild cognitive impairment and another condition contributing to CI; (4) AD: Alzheimer's disease dementia, no other condition contributing to CI (probable AD); (5) AD+: Alzheimer's disease dementia and other conditions contributing to CI (possible AD); and (6) Other dementia: other primary cause of dementia, no clinical evidence of AD. From these, 1 was defined as healthy, 2 and 3 were defined as MCI, and 4 and 5 were clubbed as AD. Ultimately this yielded our study set of 1771 individuals, out of which 1287 were healthy, 428 were MCI, and 56 belonged to the AD class.

There were 24 cognitive tests in total, out of which 19 tests could be divided into five cognitive domains, three tests were subsets of some of the individual 19 tests, and two other tests couldn't be categorized into one of the five domains. The five cognitive domains were episodic memory, perceptual orientation, perceptual speed, semantic memory, and working memory which had effectively seven, three, four, five and three tests, respectively. The two other tests were Mini-Mental State Exam (MMSE) and an auditory comprehension test. The list of the tests along with the domains is mentioned in Table S1 in supporting information.

The data had 1771 individuals, so we could easily implement seven-fold cross-validation with 253 individuals in each subset of our testing

TABLE 1 Distribution of HV, MCI, and AD across seven training and testing data subsets for ROSMAP studies

Distribution	Values			Percentage		
Data set	HV	MCI	AD	HV	MCI	AD
1	187	54	12	73.91304	21.34387	4.74308
2	181	63	9	71.5415	24.90119	3.55731
3	185	63	5	73.12253	24.90119	1.97628
4	180	61	12	71.14625	24.11067	4.74308
5	192	53	8	75.88933	20.94862	3.16206
6	179	69	5	70.75099	27.27273	1.97628
7	183	65	5	72.33202	25.6917	1.97628
Total	1287	428	56	72.67081	24.16714	3.16206

Abbreviations: AD, Alzheimer's disease; HV, healthy; MCI, mild cognitive impairment; ROSMAP, the Rush Religious Orders Study and Rush Memory and Aging Project

data. The distribution of individuals in each of these subsets was random, and the percentage of each class in the subsets can be seen in Table 1.

2.1.2 | National Alzheimer's Coordinating Center (NACC)

The NACC data set comprises data collected from approximately 30 test centers, after which the data are unified and harmonized to remove the differences. Nevertheless, data collected from several different centers introduce heterogeneity in the overall data set, and therefore we should exercise caution if we want to consider the overall data set as a single sample. So, to remove this heterogeneity, we can individually study each test center data or a combination of data from a few similar test centers would also work. The NACC data were diagnosed in two different ways. First, with a NACC-derived variable describing the cognitive status of the individual at the visit and second a NACC-derived variable described as presumptive etiologic diagnosis of the cognitive disorder—AD. In the first classification, the individual can be classified as healthy, MCI, having dementia, and cognitively impaired with no signs of MCI or dementia. From the second classification, the individual can be healthy, AD, or cognitively impaired with no AD. So, we classified healthy from both ways as healthy (and in fact the individuals who were healthy were the same from both methods), MCI were classified as those which were MCI from the first method but cognitively impaired with no AD from the second, and AD were those who were AD under the second classification and people having dementia from the first classification method. After defining the path to classify the individuals as healthy, MCI, and AD, we selected the data from test centers that can be used for our study. To maintain a well-proportioned presence of all the three classes, we adopted the following criteria. We defined the threshold that healthy volunteers should be > 45% of the total volunteers studied, the numbers of MCI individuals should be greater than ones with AD, and the representation of AD should be at least 10 individuals in each data set. With this in mind, data from only five testing centers qualified. Hence, we had 664 individuals comprising 449 healthy, 128 MCI, and 87 AD individuals. Unlike ROSMAP, the neuropsychological tests were not divided according to any cognitive domains. Thirty cognitive tests (Table S2 in supporting information) were used for our study from NACC data.

We implemented eight-fold cross-validation with 83 individuals in each subset of our testing data. The distribution of individuals in each of these subsets was random, and the percentage of each class in this subset can be seen in Table S3 in supporting information.

2.2 | Machine learning (ML) approaches

Broadly, we applied two supervised ML techniques to our data. They were random forests (RF) and support vector machine (SVM), which are widely acknowledged and accepted classification techniques. In our study, supervised learning is a better approach than the unsupervised learning because we are training a subset of data from the known classes and testing it on the remaining data along with predicting its class and supervised learning methods are known to outperform the unsupervised ML techniques in this regard. In addition, approaches like principal component analysis and factor analysis don't preserve the individual identity of the cognitive tests, which works opposite to our aim of finding the optimal combination of neuropsychological tests for better prediction accuracy.

We used packages available in the R programming language, which is a free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing[14] for executing these supervised ML techniques.

Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[15-17] Random decision forests correct for decision trees' habit of overfitting to their training set.[18] For our study, we used RF for classification purposes. In the R language using randomForest library, we had parameters, like ntree representing the number of trees to grow in the model, and mtry, which represents number of variables or predictors randomly sampled as candidates at each split. We tuned these parameters by calculating accuracy in prediction (the method used to measure accuracy is mentioned at the end of this subsection) for different testing data subsets as specified earlier. From the exhaustive combinations of ntree and mtry, which ranged from {100–1000 (in steps of 100), 2000,5000} and {1,3,6,9,12,18} respectively, we calculated the multi-class area under the curves (AUC) or receiver operating characteristic (ROC) curves to find the values of ntree and mtry which could optimize the ROC values. The multiclass ROC in R calculates the mean of the AUC for different subsets of two classes.[19] It is important to note here that while a multitude of combinations of ntree and mtry can be tested, after a certain point this leads to diminishing returns. We observed that combination of ntree as 800 and mtry as 18 for our study data led to best prediction accuracy for our RF model, as can be

seen in Figure S1 in supporting information. So, these parameters were selected for comparing RF with other ML approaches.

SVM[20] are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM can also be used as a non-linear classifier by using a different kernel. Hence, for our case we used linear kernel, which is parameterized by a linear polynomial and radial kernel, which is parameterized by a Gaussian function, and compared the accuracy in prediction from these two cases with the one obtained from the RF technique. We used both the kernels to see if our data set can be separated linearly or not into HV, MCI, and AD classes.

We calculated accuracy for prediction for different testing data subsets as specified earlier to evaluate the performance of each ML approach studied here, using function confusion matrix available in the "caret" package in R, which relies on the one versus all approach for accuracy measurements for more than two classes. The accuracy is the proportion of the total number of predictions that were correct and from this function we can also obtain the sensitivity (true positive rate or the proportion of positive cases that were correctly identified) and specificity (true negative rate or the proportion of negative cases that were identified correctly) values of each class for different models.

2.3 | Optimizing the number of neuropsychological tests

2.3.1 | Fisher's score

Fisher's linear discriminant analysis (LDA) or Fisher's score (F-score) is generalized for three classes and defined as:

$$F\ (i) = \frac{\sum_{j=1}^{3}\left(x_{avg,i}^{j} - x_{avg,i}\right)^{2}}{\sum_{j=1}^{3}\left(\frac{1}{n^{j}-1}\right)\sum_{k=1}^{n^{j}}\left(x_{k,i}^{j} - x_{avg,i}^{j}\right)^{2}}$$

where $F(i)$ represents the F-score of i^{th} feature, $x_{avg,i}^{j}, x_{avg,i}$ denote the average of i^{th} feature of j^{th} class and whole data respectively, n^{j} represents the number of j^{th} class samples, $x_{k,i}^{j}$ means the training vectors of i^{th} feature and j^{th} class. The larger the value of F-score, the more likely it is that this feature is discriminative. We used this behavior of the F-score to assign importance to the individual tests in the battery and therefore, as a preliminary method to rank the tests. The LDA formula is extended for three classes using the two-way classifier LDA used by Tan et al.[21] in our in-house script.

2.3.2 | Ranking of tests or feature selection

After obtaining the F-score, in each training data set, these tests (or features) were added one by one according to their F-score in the RF. The model was built using RF where we varied ntree from {100-1200 (in steps of 100),2000,3000,4000,5000} for three different commonly used values of mtry, that is, p/2, p/3, and p/1.414[22] where p represents the number of tests or features in that particular iteration. The minimum mtry was one for cases in which mtry on calculation from the above-mentioned formula was less than 1. The model was then tested on the testing data and we calculated the accuracy of prediction. The plots for this model are provided in Figure S2 in supporting information. This approach is reliable because we could observe all the tests or features whose addition resulted in significant rise in accuracy of prediction and therefore can be selected as an important test. From the selected important tests, we added each test iteratively according to their F-score, to find out which least combination of tests finally classified better or more similar compared to the accuracy of prediction when the whole battery of tests was used.

3 | RESULTS

3.1 | Machine learning approaches

We used ML algorithms to predict and classify the cognitive status of an individual based on their scores in a battery of neuropsychological tests by constructing training and testing data subsets. We were blinded to the final outcome of the cognitive status of the individual in the testing phase. We observed that, when averaged over different training data sets, RF models can classify individuals as HV, MCI, or AD with 95% accuracy, whereas linear and radial SVM are, respectively, 86% and 88% accurate for the same (depicted in Table 2 and Figure 1). The sensitivity of prediction for individuals with AD is modest (Table 2). Also, Figure 1 explicitly depicts that for the second subset of ROSMAP data, RF correctly predicts seven out of nine AD individuals and 52 out of 63 MCI individuals; whereas linear SVM and radial SVM correctly predict four and five AD individuals, and 44 and 42 MCI individuals, respectively. It is evident from Table 2 that overall for all the data subsets, RF is more accurate in distinguishing the healthy class from MCI or AD group, because it does not predict a healthy person as MCI or AD, whereas SVM does so, although for a few data subsets linear SVM predicts the true AD cases (sensitivity) marginally more accurately. Moreover, RF might predict AD less in number, but it predicts AD as either AD or MCI whereas SVM can easily predict AD affected individuals as healthy in most of the cases. With that said, another aspect we pondered was the follow-ups of the wrongly predicted individuals, although the current premise of this work is the baseline assessments of the study subjects. It is interesting to note that 17 MCI individuals predicted as healthy were diagnosed as healthy out of 28 such cases in the subsequent follow-ups. Moreover, six out of seven MCI individuals predicted as AD were found to have AD in the subsequent follow-ups. Furthermore, 9 out of 18 AD patients predicted as MCI were seen as MCI or healthy in subsequent follow-ups. We acknowledge that these observations could also be due to true progression of MCI to healthy or AD over a time period of 1 year. Nonetheless, this is an important

TABLE 2 Accuracy, sensitivity, and specificity of prediction models for different data sets compared over three machine learning approaches namely, random forests (ntree = 800, mtry = 18), linear and radial support vector machines for ROSMAP data

		Random forests					
		HV		MCI		AD	
Data set	Accuracy	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
1	0.9605	0.9947	0.9545	0.9444	0.9648	0.5	1
2	0.9407	0.989	0.875	0.8254	0.9789	0.77778	0.9918
3	0.9407	0.9838	0.8971	0.8413	0.9737	0.6	0.9879
4	0.9328	0.9944	0.863	0.8525	0.9635	0.41667	1
5	0.9644	1	0.918	0.8868	0.985	0.625	0.9959
6	0.9407	0.9944	0.8378	0.8261	0.9837	0.6	1
7	0.9605	0.9945	0.9143	0.8923	0.984	0.6	0.996
				Linear support vector machine			
1	0.8656	0.9358	0.7424	0.6667	0.9246	0.6667	0.9917
2	0.8538	0.9282	0.7361	0.6984	0.9105	0.44444	0.9959
3	0.8261	0.9027	0.6618	0.6032	0.9	0.8	0.9919
4	0.8617	0.95	0.726	0.6721	0.9219	0.5	1
5	0.8972	0.9583	0.7541	0.6981	0.95	0.75	0.9959
6	0.8656	0.9721	0.6486	0.6087	0.962	0.6	0.996
7	0.8538	0.9344	0.7286	0.6462	0.9255	0.6	0.9839
				Radial support vector machine			
1	0.8775	0.9572	0.7576	0.7037	0.9246	0.41667	1
2	0.8696	0.9558	0.7361	0.6667	0.9368	0.55556	0.9918
3	0.8854	0.9622	0.7059	0.6825	0.9526	0.6	1
4	0.8656	0.9722	0.7123	0.6557	0.9323	0.3333	1
5	0.8854	0.9531	0.7213	0.6792	0.94	0.625	1
6	0.8775	0.9888	0.6622	0.6377	0.9674	0.2	1
7	0.8775	0.9563	0.7286	0.6923	0.9415	0.4	0.996

Abbreviations: AD, Alzheimer's disease; HV, healthy; MCI, mild cognitive impairment; ROSMAP, the Rush Religious Orders Study and Rush Memory and Aging Project

observation that in the majority of our incorrect classifications, it aligns with the follow-up diagnosis of the individuals' cognitive status, thus implying that our visible inaccuracy in prediction might be, in part, due to the uncertainty in diagnostic procedures in the baseline visit.

We implemented the same ML approaches on NACC data set. We observed that RF, and linear and radial SVM perform similarly with an accuracy of about 77% (averaged over different training data-sets) for predicting HV, MCI, and AD individuals (Table S4 in supporting information). Despite that, we observe that RF is a better multi-class classification algorithm in this case too because RF is more sensitive in predicting MCI individuals and it very rarely predicts a healthy person as AD. Although radial SVM might predict AD individuals more sensitively, it also predicts one healthy individual as AD whereas RF doesn't predict any healthy individual as AD (refer to Table S5 in supporting information for a fourth data set contingency table of NACC data). For example, in the fourth data set, RF correctly predicts 4 MCI and 10 AD individuals whereas linear SVM and radial SVM correctly predict 3 and 1 MCI individuals, and 9 and 13 AD individuals, respectively. Both linear and radial SVM denotes MCI individuals as healthy in the classification more times compared to RF, which is not desired in such algorithms. Thus, radial SVM can delineate AD versus other two classes well and perform better as a two-way classifier, but not essentially HV, MCI, AD as three different classes (Table S5).

Multi-class classification for biomedical or other kinds of data sets is challenging, as multi-way classifiers usually perform poorly for unbalanced or high-dimensional data, whereas widely used binary classifiers, like logistic regression, require collapsing the multi-class problem into a binary design to resemble one versus one classification for ease of execution. Our method is the first report of automated multi-class classification of cognitive status based on neuropsychological tests using ML approaches in R language and we found that RF performs better for multi-class classification purposes in diverse study data sets.

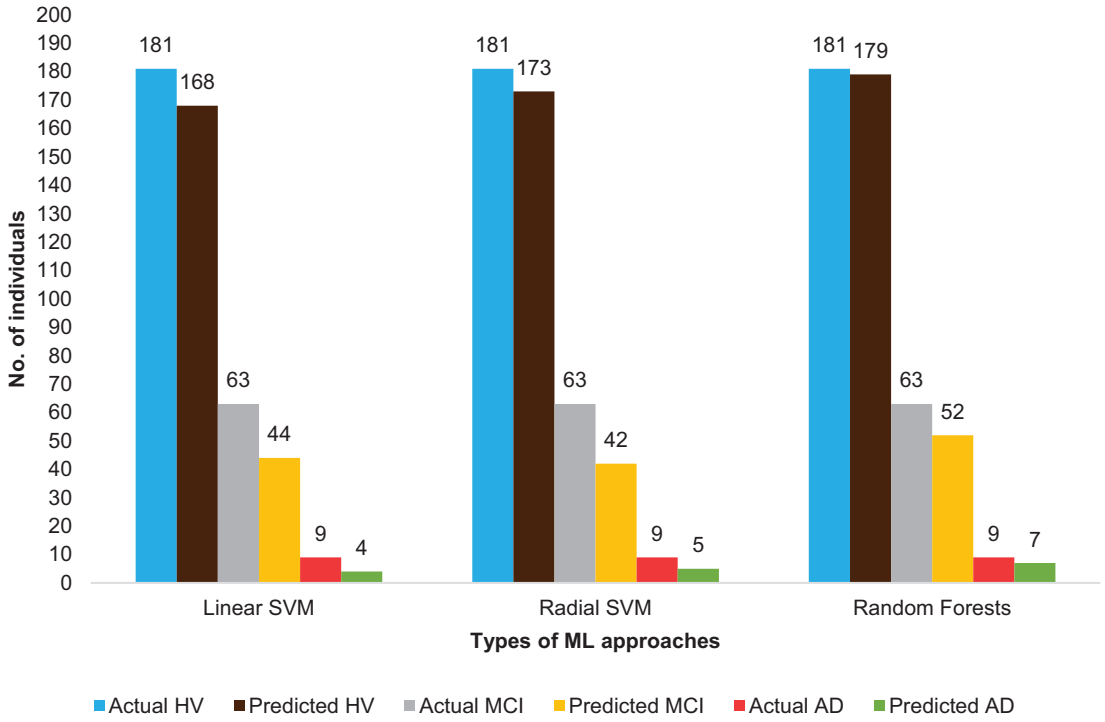

FIGURE 1 Performance comparison of three different machine learning approaches. It shows the number of actual and correctly predicted healthy (HV), mild cognitive impairment (MCI), and Alzheimer's disease (AD) individuals for the second data subset out of the seven data subsets for the Rush Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) data set. ML, machine learning; SVM, support vector machine

3.2 | Subset of decisive tests

3.2.1 | ROSMAP data

We can see the sequence of the neuropsychological tests based on their discriminating power according to their F-scores in Figure 2A. Imposing a threshold of 60% of the maximum F-score for a given study data, we found that word list recall in episodic memory domain (wlii), MMSE (mmse30), logical memory delayed recall in episodic memory (delay), category fluency in semantic memory (catflu), and logical memory immediate recall in episodic memory (story) pass this cut-off (F-scores $\geq$ 1.5). Therefore, these tests could plausibly be decisive in predicting the outcome of the entire battery when administered on an individual. Change of accuracy brought in by removing each test iteratively was also calculated to get the holistic view of ranking the tests for the data set, which can be seen in Figure 2B for ROSMAP data. For this calculation, we used the RF model with ntree = 800 and mtry as 18 based on the tuning of RF approach mentioned earlier.

Even though tests like wordlist recognition (wliii), line orientation (lopair), and nine-items progressive matrices (pmsub) were not important by Fisher's score from Figure 2B, their importance was directly reflected as their removal resulted in notably more change in accuracy compared to other tests. These tests were, thus, very crucial and removing any of them would drastically decrease the accuracy of prediction from 95% to 90% if either one of the tests, for example, line orientation or progressive matrices (subset), were removed (Figure S3 in supporting information). Combining the above two approaches, we note that the eight tests: wlii, mmse30, delay, wliii, pmsub, lopair, catflu, story could accurately predict and classify the cognitive status of an individual as HV, MCI, or AD. Out of these eight tests, four belonged to episodic memory, namely word list recall, word list recognition, logical memory–immediate, and delayed recall; two belonged to perceptual orientation, namely line orientation and subset of progressive matrices; one belonged to semantic memory, namely category fluency; and one was MMSE.

Existing feature selection methods may seem handy; however, they work on preliminary elimination and final variable selection where features eliminated from the first step are not used for the second step and unlike our approach as we used preliminary ranking via F-score and then assessed the change in accuracy by removing each test iteratively where we included all the tests. This is crucial because eliminating a subset of tests based on their low importance using just a single statistical approach may lead to loss of important variables, for example, while implementing an existing feature selection method VSURF[23] on the ROSMAP data, we found that line orientation never turned out to be important for any of the training and testing data, but we clearly observe from our approach that this test is significant for accurate classification of the cognitive status. Moreover, our method is designed such that each step is properly understood while being implemented, which also adds to why we didn't use any existing feature selection methods.

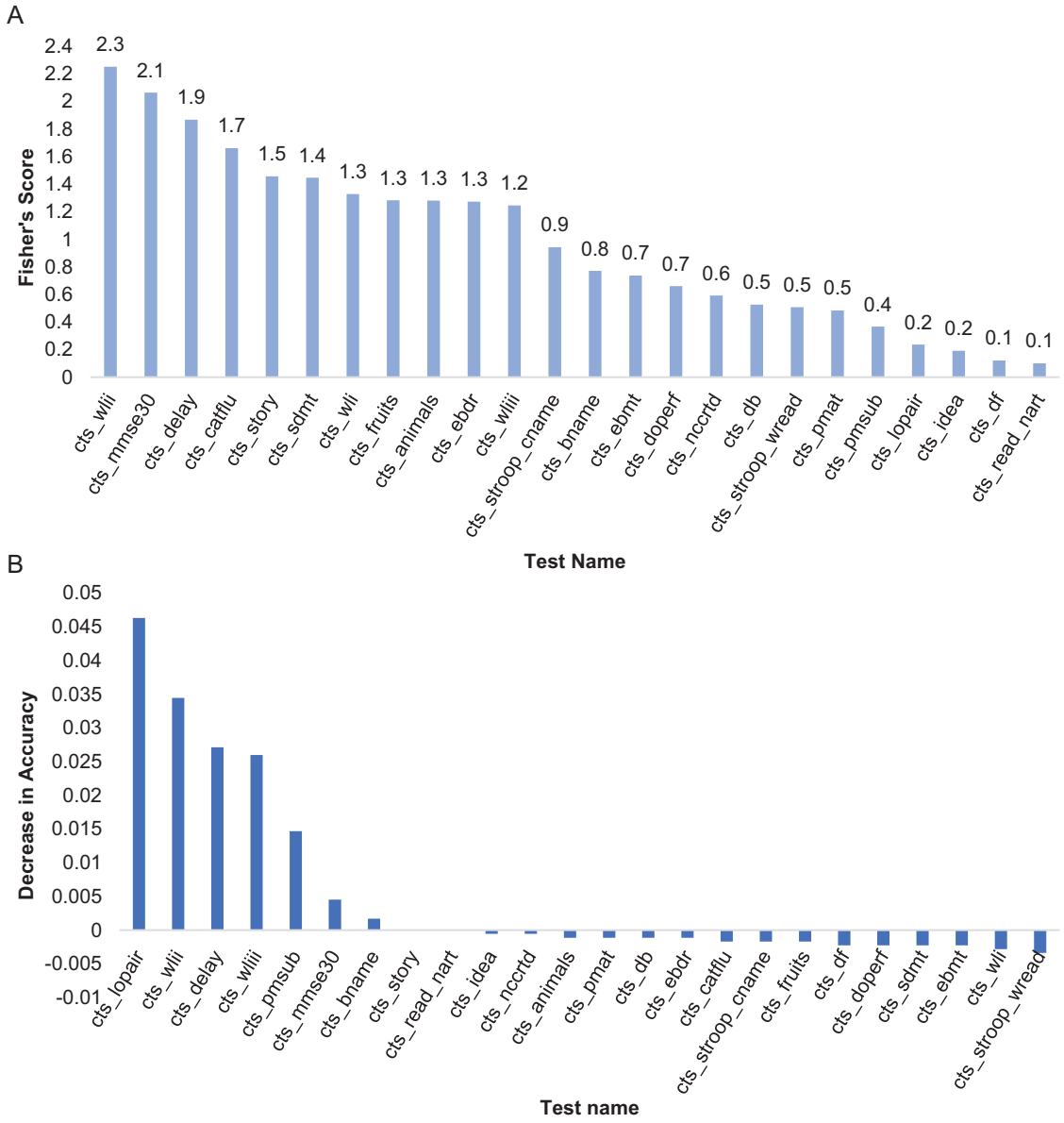

FIGURE 2 Ranking of 24 neuropsychological tests from the Rush Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) study. A, Fisher's score for three classes classification. B, Average decrease in accuracy of prediction when a single test was removed iteratively

Using the feature selection methods mentioned earlier, by which we used RF with ntree = 800 and mtry as p/2 where p represents the number of tests or variables in that particular iteration, we found that average accuracy of prediction for eight cognitive tests was roughly 95% whereas for the whole battery of tests it was 94%. This slight increase in prediction accuracy might be due to the fact that some cognitive tests are redundant and some might not be a good measure for a multiclass prediction. Figure 3 depicts the actual and correctly predicted numbers of HV, MCI, and AD individuals for three subsets of data. For example, in the second data subset in which only eight cognitive tests were chosen, the number of correctly predicted HV, MCI, and AD individuals were 178, 54 and 7 out of 181, 63 and 8 individuals, respectively, whereas the whole battery of tests correctly predicted 177 HV, 51 MCI, and 7 AD individuals, which also highlights the slight increase in accuracy of prediction for subset of tests. We show that the accuracy of prediction for two other combinations of tests (six from the above eight, nine tests comprising symbol digits modality oral test in perceptual speed domain [sdmt] along with the above eight) are comparable to that of the earlier mentioned eight tests (Table S6 in supporting information), thus again highlighting the fact that these tests are good predictors for cognitive status of an individual and accuracy of prediction is equivalent to the entire battery of neuropsychological tests.

3.2.2 | NACC data

We find that delayed recall verbatim (CRAFTDRE), delayed recall paraphrase (CRAFTDVR), Montreal Cognitive Assessment (MoCA), delayed

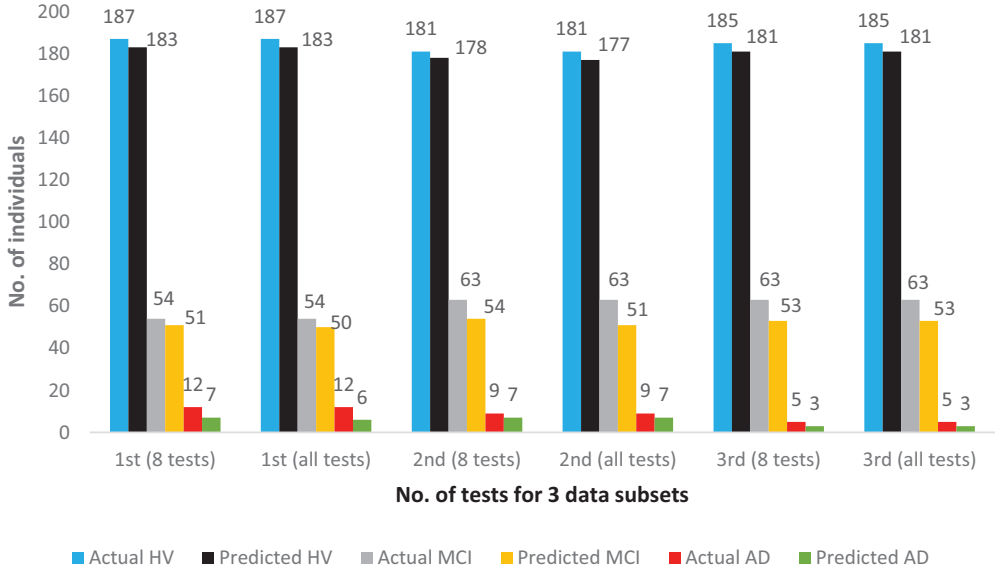

FIGURE 3 Performance comparison of subset of eight tests to the whole battery for the Rush Religious Orders Study (ROS) and Rush Memory and Aging Project (MAP) study. It shows the number of actual and correctly predicted healthy (HV), mild cognitive impairment (MCI), and Alzheimer's disease (AD) individuals for three testing data subsets comparing results for eight tests versus all 24 tests using random forests (ntree = 800 and mtry = [number of tests/two] meaning 4 and 12, respectively)

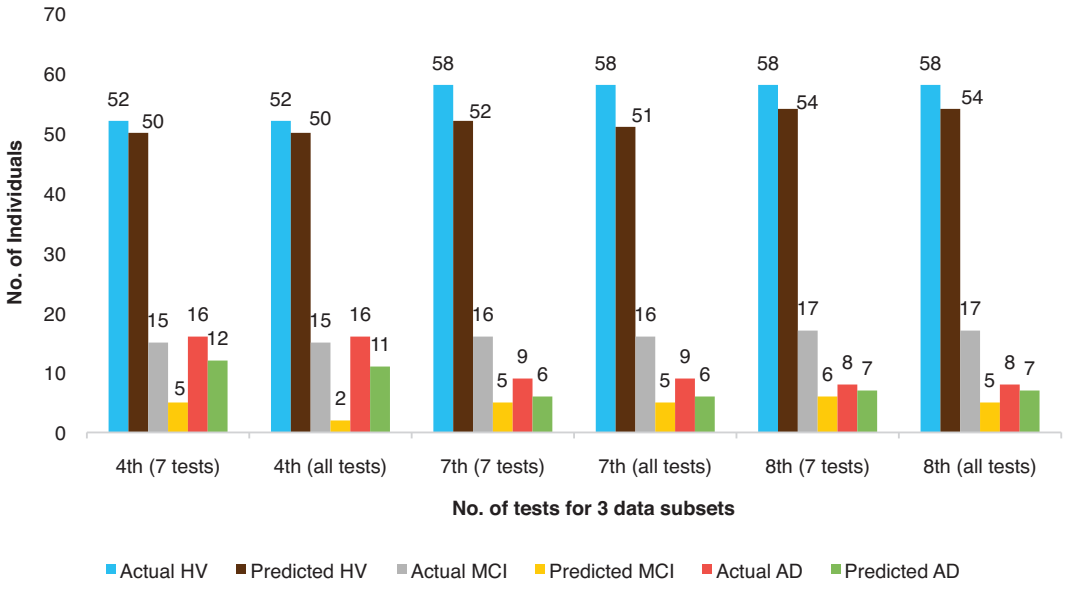

FIGURE 4 Performance comparison of subset of seven tests to the whole battery for National Alzheimer's Coordinating Center study. It shows the number of actual and correctly predicted healthy (HV), mild cognitive impairment (MCI), and Alzheimer's disease (AD) individuals for three testing data subsets (which are fourth, seventh, and eighth data sets) comparing results for 7 tests versus all 30 tests using random forests (ntree = 800 and mtry = [number of tests/two] meaning 4 and 15, respectively). ML, machine learning; SVM, support vector machine

copy of Benson figure (UDSBENTD), vegetable naming in 60 seconds (VEG), trail making test (TRAILB), immediate recall verbatim (CRAFTVRS), and immediate recall paraphrase (CRAFTURS) have F-score > 0.8, which is greater compared to other tests which have F-score < 0.6. Therefore, by preliminary F-score evaluation (roughly 60% of the maximum F-score of the given study data) we have these tests important as depicted in Figure S4A in supporting information. Further performing our feature selection method, we obtained seven important tests (excluding the CRAFTVRS in the above mentioned eight tests) which yielded average prediction accuracy of 78% compared to 77% for that of the whole battery of tests as shown in Table S7 in supporting information. Figure 4 depicts the numbers of actual and

correctly predicted HV, MCI, and AD individuals for three data subsets (fourth, seventh, eighth subsets) for NACC revealing how the combination of seven tests can be slightly better in prediction compared to the whole battery; for example, in the fourth data subset, 5 and 12 individuals are correctly predicted as MCI and AD, respectively, while including only these seven tests whereas the whole battery of tests correctly predicted only 2 MCI and 11 AD individuals. These seven important tests were a part of the top eight tests according to the F-score. Here, F-score played a pivotal role instead of the change in accuracy of prediction by removing one test iteratively (unlike the case of ROSMAP data), which was calculated here too and can be seen in Figure S4B. This is due to the fact that NACC data are considerably heterogeneous as they tried to unify different tests and data from different disease centers. In spite of the fact that the study data are heterogeneous, which affected the overall accuracy in prediction value, our feature selection approach performed well in identifying the important cognitive tests. Therefore, our method of feature selection discerns the combination of tests that could be decisive in predicting and classifying the cognitive status of individuals into HV, MCI, or having AD, in population study as well as with heterogeneous data from various AD clinics.

4 | DISCUSSIONS

4.1 | Conclusions

We have described in this work that RF performs better for prediction and classification for our multi-class cognitive data set compared to that of SVM with linear or radial kernel. Also, we can highlight from our findings that a combination of eight tests can classify individuals as HV, MCI, and AD with as much accuracy as using the whole battery of 24 neuropsychological tests in the ROSMAP data and combination of seven tests performs as good as all the 30 neuropsychological tests taken together for NACC. It is striking to note here that the type of tests which are important are similar and belong to the same domains in both study data sets. They are not exact as they are performed by different population studies and disease testing centers. For example, comparing both the study sets we find that MMSE in ROSMAP is similar to MoCA in NACC; wlii, delay, and wliii are delayed recall tests in ROSMAP like CRAFTDVR, and CRAFTDRE in NACC. Similarly, pmsub and lopair from ROSMAP are pattern making and line orientation tests, respectively, which can be related to UDSBENTD and TRAILB from NACC which are, respectively, delayed copying of Benson figure and trail making. Also, catflu (category fluency) of ROSMAP is like VEG (vegetable naming) of NACC, and story, which is immediate story recall in ROSMAP is equivalent to CRAFTURS in NACC. MMSE or MoCA are very crucial tests in distinguishing AD from non-AD individuals and are therefore very important in any kind of cognitive study. Our work shows that, in conjunction with the other six or seven tests from the neuropsychological battery, they can accurately classify individuals as HV, MCI, or AD.

In previous such studies, the importance of delayed and immediate recall along with some other features, and subjective guidance by neuropsychologists have been highlighted in predicting AD individuals.[10] Here we have more rigorously shown with multiple data sets how a subset of eight neuropsychological tests could be decidedly used for predicting individuals as HV, MCI, and AD, in a fully automated ML model.

4.2 | Future actions and description

As a generalizable application, we would encourage researchers to adopt our feature selection approach for testing and implementation in other feature selection and classification studies. This is because our method involves a ranking approach combined with brute-force examination of the variables involved in the study data which is worth testing in other types of data sets in addition to neuropsychological batteries.

Coming back to the field of AD research, it would be useful to test our model on other AD study data sets or population studies aimed at understanding cognitive decline. These results are promising, and we look forward to its large-scale implementation in communities. The National Institute on Aging–Alzheimer's Association (NIA-AA) framework treats cognitive impairment as a symptom/sign of AD rather than the definition of the disease,[24] but it is still the most prominent, effective, inexpensive, and preliminary way to diagnose cognitive impairment as a part of AD. However, more often than not, the neuropsychological batteries are time-consuming and exhausting for the study subjects. Implementation of our results will be beneficial in reducing time taken to administer the tests while ensuring minimal exhaustion and burden for the study individuals in addition to providing cost-effective measures for conducting large-scale research studies tailored to classify individuals as HV, MCI, or AD. Every cohort can begin with the whole battery of tests, and subsequently, this feature selection model can be implemented which will identify a subset of neuropsychological tests for that particular study group and hence facilitate expeditious and accurate classification of cognitive status in research and community studies. We would also like to propose an extended application of our results for neuropsychological tests when they could be combined with AD pathology biomarkers in the above-mentioned ML models and subsequently tested for accuracy of prediction of cognitive status. This could help maximize the accuracy of prediction and classification and obtain the best predictors for the population.

ACKNOWLEDGMENTS

We thank Prof. Vijayalakshmi Ravindranath (Director, Centre for Brain Research) and Prof. Naren Rao (Additional Professor of Psychiatry, National Institute of Mental Health and Neuro Sciences) for the critical discussions in the preliminary stages of this work. This research work was supported by CBR startup funding to BK. The research for ROSMAP (data set used in this work) was supported by NIH grants P30AG010161, R01AG015819, and R01AG17917. We thank Dr. David Bennett (PI) and Rush Alzheimer's Disease Center for ROSMAP data access. The NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded

ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30 AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & dementia*. 2013;9(1):63-75
2. Nichols E, Szoeke CE, Vollset SE, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*. 2019;18(1):88-106
3. Prince M, Wimo A, Guerchet M, Ali GC, Wu YT, Prina M. *The Global Impact of Dementia. An Analysis of Prevalence, Incidence, Cost and Trends*. London: Alzheimer's Disease International (ADI); 2015
4. Querfurth HW, LaFerla FM. Mechanisms of disease. *N Engl J Med*. 2010;362(4):329-344
5. Etters L, Goodall D, Harrison BE. Caregiver burden among dementia patient caregivers: a review of the literature. *J Am Acad Nurse Pract*. 2008;20(8):423-428
6. Lezak MD, Howieson DB, Loring DW, Fischer JS. *Neuropsychological Assessment*. USA: Oxford University Press; 2004
7. Bennett DA, Schneider JA, Aggarwal NT, et al. Decision rules guiding the clinical diagnosis of Alzheimer's disease in two community-based cohort studies compared to standard practice in a clinic-based cohort study. *Neuroepidemiology*. 2006;27(3):169-176
8. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med*. 2001;23(1):89-109
9. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf Fusion*. 2019;50:71-91
10. Battista P, Salvatore C, Castiglioni I. Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study. *Behav Neurol*. 2017;2017
11. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. *Curr Alzheimer Res*. 2012;9(6):628-645
12. Rush Alzheimer's Disease Centre -RADC_codebook_data_set_603_12-07-2017 https://www.radc.rush.edu/ (Last visited on September 20, 2019)
13. NACC UDS Researcher's Data Dictionary (Version 3.0, March 2015). National Alzheimer's Coordinating Center (Walter A. Kukull, PhD, Director), University of Washington, USA. https://www.alz.washington.edu/WEB/forms_uds.html Assessed on 28 August 2019
14. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (Last visited on September 20, 2019)
15. Ho TK. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition. *IEEE*. 1995;1:278-282
16. Barandiaran I. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20(8):1-22
17. Cutler DR, Edwards Jr TC, Beard KH, et al. Random forests for classification in ecology. *Ecology*. 2007;88(11):2783-2792
18. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*. 2005;27(2):83-85
19. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn*. 2001;45(2):171-186
20. Cortes C, Vapnik V. Support-vector networks. *Machine learning*. 1995;20(3):273-297
21. Tan NX, Rao HB, Li ZR, Li XY. Prediction of chemical carcinogenicity by machine learning approaches. *SAR QSAR Environ Res*. 2009;20(1-2):27-75
22. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18-22
23. Genuer R, Poggi JM, Tuleau-Malot C. VSURF: an R package for variable selection using random forests. *The R Journal*. 2015;7(2):19-33
24. Jack Jr CR, Bennett DA, Blennow K, et al. NIA-AA Research Framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*. 2018;14(4):535-562

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Gupta A, Kahali B. Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. *Alzheimer's Dement*. 2020;6:e12049. https://doi.org/10.1002/trc2.12049