Received: 9 June 2023 | Revised: 19 September 2023 | Accepted: 26 September 2023

DOI: 10.1002/dad2.12494

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Improving clinical efficiency in screening for cognitive impairment due to Alzheimer's

Yueqi Ren[1,2] | Babak Shahbaba[1,3] | Craig E. L. Stark[1,4]

[1]Mathematical, Computational and Systems Biology Graduate Program, Center for Complex Biological Systems, University of California Irvine, Irvine, California, USA

[2]Medical Scientist Training Program, School of Medicine, University of California Irvine, Irvine, California, USA

[3]Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, California, USA

[4]Department of Neurobiology and Behavior, University of California Irvine, Neurobiology and Behavior, Irvine, California, USA

Correspondence
Craig E. L. Stark, Mathematical, Computational, and Systems Biology Graduate Program, Center for Complex Biological Systems, University of California Irvine, 1424 Biological Sciences III, Irvine, CA 92697, USA.
Email: cestark@uci.edu

Funding information
National Institute on Aging, Grant/Award Numbers: P30AG066519, F30AG079610; National Institute of Mental Health, Grant/Award Number: R01MH115697

Abstract

INTRODUCTION: To reduce demands on expert time and improve clinical efficiency, we developed a framework to evaluate whether inexpensive, accessible data could accurately classify Alzheimer's disease (AD) clinical diagnosis and predict the likelihood of progression.

METHODS: We stratified relevant data into three tiers: obtainable at primary care (low-cost), mostly available at specialty visits (medium-cost), and research-only (high-cost). We trained several machine learning models, including a hierarchical model, an ensemble model, and a clustering model, to distinguish between diagnoses of cognitively unimpaired, mild cognitive impairment, and dementia due to AD.

RESULTS: All models showed viable classification, but the hierarchical and ensemble models outperformed the conventional model. Classifier "error" was predictive of progression rates, and cluster membership identified subgroups with high and low risk of progression within 1.5 to 3 years.

DISCUSSION: Accessible, inexpensive clinical data can be used to guide AD diagnosis and are predictive of current and future disease states.

KEYWORDS
Alzheimer's disease, clinical efficiency, conversion rate, diagnosis classification, prediction, progression monitoring, screening, statistical machine learning

HIGHLIGHTS

- Classification performance using cost-effective features was accurate and robust
- Hierarchical classification outperformed conventional multinomial classification
- Classification labels indicated significant changes in conversion risk at follow-up
- A clustering-classification method identified subgroups at high risk of decline

1 | BACKGROUND

Projected to reach 13 million by 2050, the high number of patients with Alzheimer's disease (AD) is contrasted by a shortage of clinical experts in the United States, with a shortfall of two-thirds of the number of geriatricians needed by 2050.[1] Considering these trends, improving clinical efficiency and empowering primary care providers to screen patients for AD are paramount goals. A recent meta-analysis of dementia diagnoses made by general practitioners found that distinguishing between patients with and without dementia produced an average F1-score of 0.735, with high levels of heterogeneity across studies.[2] These findings illustrate the existing gap in screening for AD in primary care settings. Achieving reliable screening at primary care visits would enable better allocation of clinical resources and improve access to specialty services.

Few studies in the field of AD classification have focused on improving clinical efficiency in a primary care setting. Most studies in this area have either utilized multimodal biomarkers, including neuropsychological testing, MRI, PET imaging, and cerebrospinal fluid (CSF) biomarkers,[3–5] or extensively focused on specific modalities of data.[6–10] Many studies have also included predictive features from the Clinical Dementia Rating (CDR) Dementia Staging Instrument.[4,9,11,12] Despite their popularity in the literature, these data modalities are costly and not available for most of the aging population. In contrast, a limited number of studies have focused on cost-effective data for AD screening. One such study utilized widely accessible data to train machine learning classifiers to distinguish between cognitively unimpaired and impaired older adults.[13] Another study used survey data to perform unsupervised clustering to identify different subgroups at risk of developing dementia.[14] To our knowledge, no study has performed multinomial classification of different AD clinical diagnoses using widely accessible data.

To address this, we propose to focus on lower-cost, readily accessible data in developing statistical machine learning tools to aid clinical decision-making. These data include patient history, medical history, simple cognitive testing, and behavioral survey information that are less time-consuming and lower cost to obtain for the older adult population. We aim to provide a measure of confidence associated with the classifications. Without this, outputs from machine learning algorithms are difficult to understand and apply clinically.[15] We use longitudinal data to determine the clinical trajectory of diagnoses as a proxy for classification confidence. While past work demonstrated that classifiers, when given data at multiple time points, could predict longitudinal diagnoses, few to none have developed classification methods not exposed to information from follow-up visits that can give insight into conversion risk as a proxy for classification confidence.[8,12,16,17] In addition, it is unclear how well classifiers can leverage cost-effective data from a single visit to inform us of conversion risk at later time points. We compare several classification methods, as past work demonstrated the viability of different machine learning classifiers, including ensemble classifiers, in predicting AD status and cognitive changes.[9,18,19] We also present a novel classification method to contrast against existing methods to determine the optimal prediction performance using lower-cost features. We hypothesize that utilizing a combination of classification and clustering methods, which have been demonstrated to be useful in identifying meaningful clinical subgroups in AD,[14,20] will allow us to stratify subjects into subgroups that may be of interest for clinical intervention.

In this study, we addressed two major questions related to improving clinical efficiency using cost-effective data. First, how well can we classify clinical diagnoses made by experts using cost-effective data? Second, can cost-effective data inform us of the risk of subsequent decline? Multinomial classification results from our study using lower-cost data outperformed findings in the literature using higher-cost data and more complex deep learning classifiers.[4,5] We observed that the classifications were significantly leading the clinical diagnosis at follow-up visits within 1.5 and 3 years. We also discovered high-risk subgroups of subjects that had significantly elevated conversion rates when compared to the population rate. Thus, using only lower-cost data can enable robust classification of clinical diagnosis and simultaneously inform primary care providers of an individual's susceptibility for progressing to a more severe diagnosis at follow-up. Taken together, these important pieces of clinical information can empower care providers to refer high-risk patients to specialty care and improve access to dementia care for older adults.

RESEARCH IN CONTEXT

1. **Systematic review**: Using PubMed, the authors conducted a literature review for machine learning in classifying AD diagnosis. Most studies using multinomial classification focus on neuroimaging and extended neuropsychological testing modalities that are not widely accessible. There is a need to improve clinical efficiency by exploring AD classification using cost-effective features.
2. **Interpretation**: Our proposed methodology is a robust and generalizable framework for multinomial classification of AD diagnosis. We developed novel methods, including hierarchical and clustering-based classification, which performed comparably to and better than existing methods. We conducted analyses to reveal significant trends for progression risk at follow-up using only cost-effective data from baseline. Extending upon past work, we identified a set of high-performing, lower-cost features to improve access to AD screening.
3. **Future directions**: We encourage future studies to replicate our methods on independent datasets and incorporate plasma biomarkers as an additional modality.

TABLE 1 Summary demographics of study dataset.

Feature		Visit 1 (*N* = 1918)	Visit 2 (*N* = 1020)	Visit 3 (*N* = 353)
Mean age (SD, years)		76 (7.16)	77 (6.96)	78 (7.52)
Sex (% female)		57.2	56.2	57.5
Mean education level (SD, years)		15 (3.58)	16 (3.14)	16 (2.93)
Race (%)	White	84.0	83.1	79.9
	Black or African American	10.8	11.3	12.5
	Asian	2.19	1.86	3.40
	Other	3.01	3.14	4.2
Ethnicity: Hispanic or Latino (%)		10.1	8.04	8.78
Diagnosis (%)	Cognitively unimpaired (HC)	1129 (58.8%)	610 (59.8%)	240 (68.0%)
	Mild cognitive impairment (MCI)	347 (18.1%)	130 (12.7%)	49 (13.9%)
	Dementia due to AD	442 (23.0%)	280 (27.5%)	64 (18.1%)

Abbreviations: AD, Alzheimer's disease; HC, healthy controls; SD, standard deviation.

2 | METHODS

2.1 | Subjects

The data utilized in this study were obtained from the National Alzheimer's Coordinating Center (NACC).[21] NACC standardizes data collected across more than 30 Alzheimer's Disease Research Centers (ADRCs) in the United States. The Uniform Data Set (UDS) contains clinical information, behavioral survey responses, neuropsychological testing results, and additional diagnostic information for each subject.[22,23] This study used data from 17 ADRCs (anonymized to protect privacy), including UDS visits from September 2005 to December 2020. These ADRCs are located at major academic and research institutions in the United States, and some have specific research and recruitment focuses that serve to advance dementia research. Some ADRCs focus on recruiting underrepresented populations in dementia research to improve the generalizability of findings from these data. All subjects within this analysis had at least one UDS visit with an associated magnetic resonance imaging (MRI) scan within 1 year of the UDS visit. At each UDS visit, subjects received a diagnosis of either cognitively unimpaired (healthy controls, HC), mild cognitive impairment (MCI) due to AD, or dementia due to AD. We identified the first UDS visit for each subject in this dataset, and for subsequent analyses the second UDS visits were defined as the next visit within 1.5 years after the first visit. Similarly, the third visits were defined as the next visit within 1.5 years after the second visit (within 3 years of the first visit). Features from the UDS and MRI modules of the NACC dataset were included in this study, with features stratified into different tiers based upon clinical experts' opinions on the accessibility and costs associated with collecting each feature. Because our focus was on the more common late-onset AD, we focused on subjects older than the age of 65 at their first visit. In total, the dataset included 1918 subjects with visit 1 data, 1020 subjects with visits 1 and 2 data, and 353 subjects with visits 1 through 3 data. Summary demographic information for subjects included in this study is presented in Table 1.

2.2 | Data processing

To account for differences in the scaling of input data, we normalized all continuous features, one-hot encoded all discrete features, and ordinally encoded all ordinal features. We discarded any features with more than 20% missing values and imputed any remaining missing data using simple value imputation. Due to the change in the neuropsychological testing battery introduced in UDS version 3.0 in 2015, we converted the newer variables to the older variables using the Crosswalk Study, which is the only systematic comparison of these two testing batteries, to our knowledge.[24,25] Structural MRI features were uniformly preprocessed by the NACC MRI processing pipeline.[26] From this we obtained processed summary volumetric and cortical thickness measures, including the volume of white matter hyperintensities. Additional methodological details can be found in the Supplemental Methods section.

2.3 | Cost tiers

A central tenet of our approach is that not all data features are equally accessible or easily obtained. After consulting with several neurologists and neuropsychologists affiliated with the University of California, Irvine ADRC (who routinely interview older adults with and without AD), we have devised a three-tiered system to categorize features based on the cost required to obtain the associated information.

Tier 1 features are "lower cost" and may be obtainable at primary care visits by a general practitioner. These features are derived from the UDS to represent information from patient interviews, medical history details, and routine screening questions asked at a primary care visit (e.g., using the Geriatric Depression Scale in the UDS to proxy the Patient Health Questionnaire-2 commonly used in practice). For older adults, especially those with memory complaints, the Mini-Mental State Examination (MMSE) is often used as an additional screening tool at primary care visits. While not sufficient to diagnose AD or related dementias alone, we included the MMSE here because it is currently one of the most used screening tools at primary care visits.[27] The scope of these features are similar to the information obtained from Medicare's Annual Wellness Visit for older adults, which has been shown to be useful for reducing the prevalence of delayed dementia diagnoses.[28] To our knowledge, data collected in this dataset were administered in English, which is a limitation for generalizing to other populations, but these features are relatively simple to administer and can be readily obtained across different older adult populations, including different languages and cultures.

Tier 1 features included demographic information (age, sex, race and ethnicity, primary language, marital status, education level), patient history information (living situation, family history, medications, health history), physical exam information (height, weight, blood pressure, heart rate, any focal neurological signs), questionnaires including the Geriatric Depression Scale (GDS), NACC Functional Assessment Scale (FAS), and Neuropsychiatric Inventory Questionnaire (NPI-Q), and MMSE. We included the behavioral survey data here because these features from the UDS are the closest proxies of the patient interview information that could reasonably be obtained from a primary care visit.

Tier 2 features are "medium cost" and primarily obtained at specialty care visits. Although it is possible some might be adaptable to primary care settings, these features require more expert involvement and patient time than tier 1 data. Furthermore, these features are more difficult to translate across languages and cultures, as some specialized tests have not been validated across different older adult populations. Tier 2 features included additional neuropsychological testing features beyond the MMSE and also included MRI features. Structural MRI requires additional resources beyond a clinical visit, which justifies its inclusion in tier 2. In summary, tier 2 features included categorical fluency tests (vegetable and animal naming), the Trails Making Test A and B, the Boston Naming Test (BNT), Logical Memory II—Delayed, digit span tests, and MRI features.

Tier 3 features are "higher cost" and are primarily used for research purposes. These included the CDR scores and apolipoprotein E (APOE) $\varepsilon 4$ carrier status from genotyping. Obtaining CDR scores typically requires consensus from a clinical team and takes significant amounts of expert time. Genotyping is generally not recommended currently for routine evaluation of patients with late-onset AD.[29]

Our three-tiered, cost-based feature stratification system enabled us to ask specific research questions about the tradeoff between information usefulness (measured by classification results) and information accessibility (measured by the cost tier). A summary of the modalities of features included in each tier is presented in Table 2.

2.4 | Classification pipeline

To build reliable classification models for differentiating between clinical diagnoses, we developed complete classification pipelines for each model of interest. To determine how well classifiers performed using lower-cost data, we selected several classifiers to compare from the existing literature, in addition to two novel approaches we developed in this study. The pipelines all consisted of a class balancing step to account for the larger prevalence of cognitively unimpaired individuals in the dataset. Next, the pipelines included a feature selection step to reduce the dimensionality of the input features, which helped to prevent overfitting to the training data and improved generalization to unseen data. Finally, the classifier was trained on a portion of the data and evaluated on a held-out set of subjects. The Synthetic Minority Oversampling Technique (SMOTE), with support for categorical features, was used to correct for the class imbalance.[30] We compared pipelines with and without this sample balancing step and found that pipelines with balancing yielded better classification results. Several dimensionality reduction and feature selection methods, including principal component analysis,[31] LASSO,[32] Bayesian ridge regression,[33] and extra trees,[34] were initially compared. Classification performance did not change significantly across dimensionality reduction methods, and to allow for flexible feature selection to include nonlinear associations, we decided to use extra trees to select for features that have feature importance values above the mean.

We used nested cross-validation (CV) to train and evaluate all pipelines. Implementation details are included in the Supplemental Methods section. We reported mean model performance across the CV folds using classification accuracy and F1-score, a summary statistic of the model's specificity and sensitivity. These raw performance metrics, which are influenced by diagnostic class prevalence, were reported alongside balanced performance metrics, which average across metrics within each class to equally weigh each diagnostic category. We calculated the classification metrics using a baseline, naive classifier that made predictions based solely on the most prevalent class. This provided us with a chance rate against which we could compare our metrics. In building our pipelines, we used the scikit-learn and imbalanced-learn packages in Python.[35,36]

2.5 | Existing machine learning classifiers

We chose to implement random forest (RF) classifiers for our study, primarily due to their proven efficacy in previous work on Alzheimer's disease (AD) and their advantageous features, such as the ability to incorporate multiple types of features, handle complex nonlinear interactions between features, and provide interpretability of important features.[4,9,37] Previous studies also demonstrated that ensemble methods can achieve high performance in AD classification.[19,38] This motivated us to utilize ensemble methods and implement auto-sklearn to automate the process of hyperparameter tuning and model selection for our study.[39] Auto-sklearn enabled us to simultaneously compare multiple different machine learning models, including gradient boosting, extra trees, support vector classifiers, multilayer perceptron, and RF, as well as evaluate both individual and ensemble model performances. Auto-sklearn has several built-in methods for sample balancing and feature selection that can be tuned during model training. We implemented this in lieu of the SMOTE and extra trees methods described earlier.

TABLE 2 Modalities and features used for analysis stratified by cost tier.

Tier	Modality (*N*)	Features
1: Lower cost (primary care)	Demographics (10)	Age, sex, education level
		Race, ethnicity, language(s) spoken
		Marital status, living situation
	Patient history (71)	Tobacco use, alcohol use, medications
		Cardiovascular conditions and comorbidities (eg, congestive heart failure, stroke, diabetes)
		Family history of dementia
		Physical exam (eg, heart rate, blood pressure, BMI)
	Behavioral surveys (39)	NACC Functional Assessment Scale
		Neuropsychiatric Inventory Questionnaire
		Geriatric Depression Scale
	Neuropsychological testing (1)	Mini-Mental State Exam
2: Medium cost (specialty)	Neuropsychological testing (12)	Logical Memory II—Delayed, Trials A and B, Boston Naming Test, vegetable and animal naming, digit span
	MRI (155)	Gross volumes, regional volumes, and cortical thicknesses
3: Higher cost (research)	Genetic testing (1)	APOE allele carrier status
	CDR (8)	CDR global, CDR sum of boxes, subdomain scores

Abbreviations: CDR, Clinical Dementia Rating; NACC, National Alzheimer's Coordinating Center.

2.6 | Hierarchical classifiers

To contrast against existing classification methods, we developed three related hierarchical classification models to leverage the inherent hierarchical structure among the three diagnostic classes (HC vs. MCI vs. AD dementia). We formulated a two-step approach by breaking down the multiclass landscape into simpler and more manageable binary classification problems (Figure 1). To this end, we explored two strategies based on first identifying those who had dementia (dementia-first model) or those who were cognitively unimpaired (unimpaired-first model). In the dementia-first model, the first level of the hierarchy identified individuals with AD dementia and the second level of the hierarchy only contained individuals who are classified as non-demented. For these subjects, we implemented a second classifier that distinguished between individuals who were cognitively unimpaired and who had MCI. For the unimpaired-first model, the first level identified individuals without cognitive impairment and the second level of hierarchy then distinguished between MCI and AD dementia among subjects classified as cognitively impaired. We used separate RF classifiers for each layer of the hierarchy.

In addition, we combined the predictions from these two hierarchical models in an ensemble that algorithmically determined the predicted diagnoses (combined model). The combined model first prioritized predictions of AD dementia from the unimpaired-first model, followed by predictions of HC from the dementia-first model. If neither of these predictions was made, the combined model then prioritized predictions of HC from the unimpaired-first model, followed by predictions of AD dementia from the dementia-first model. Predictions of MCI were incorporated at the end if no other class prediction was made. In essence, the combined model fused the classifications provided by the dementia-first and unimpaired-first models in a deterministic manner.

2.7 | Subgroup identification

Past studies incorporated multiple data modalities, including a higher-cost neuropsychological testing battery and MRI features, to generate informative clusters that reflect the heightened risk of worsening cognitive status over time.[20,40] These studies explored a variety of clustering methods, including multilayer clustering, k-means clustering, and mixture model clustering, to identify clinically meaningful subgroups in AD datasets.[41] Here, we implemented a novel clustering-based classification method to derive useful subgroup information and prediction performance metrics simultaneously. We chose to implement Gaussian mixture model (GMM) clustering for its ease of implementation, generalizability to common cluster shapes, and probabilistic cluster membership interpretation. GMM clustering can also be interpreted as soft k-means clustering due to its probabilistic cluster assignments. We applied extra trees as the feature selection step prior to clustering to reduce the dimensionality of the feature space and allow for better cluster quality. Since the number of clusters is a hyperparameter chosen a priori, we compared the algorithm's behavior using different numbers of clusters by performing silhouette analysis.[42] A cluster size of five was chosen to balance both the quantity of clusters to identify potentially meaningful subgroups and the quality of clusters.

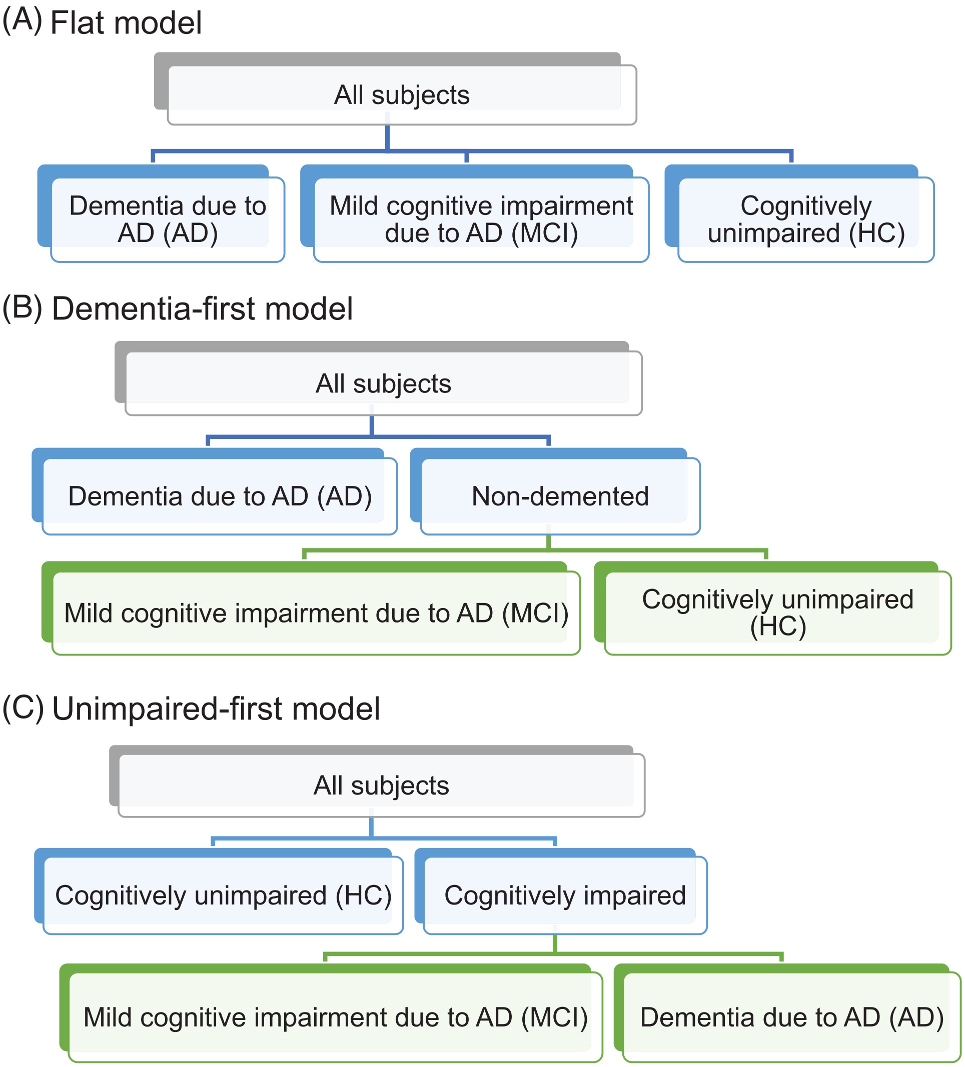
(A) Flat model
All subjects
Dementia due to AD (AD)
Mild cognitive impairment due to AD (MCI)
Cognitively unimpaired (HC)
(B) Dementia-first model
All subjects
Dementia due to AD (AD)
Non-demented
Mild cognitive impairment due to AD (MCI)
Cognitively unimpaired (HC)
(C) Unimpaired-first model
All subjects
Cognitively unimpaired (HC)
Cognitively impaired
Mild cognitive impairment due to AD (MCI)
Dementia due to AD (AD)

FIGURE 1 Hierarchical classification strategies. (A) The conventional multiclass classification setup is "flat" without any hierarchical relationships between classes. (B) The dementia-first model involves identifying subjects with AD dementia on the first level, then distinguishing between the remaining subjects. (C) The unimpaired-first model first identifies cognitively unimpaired subjects, then distinguishes between the MCI and AD dementia classes.

In the context of classification, we used five-fold CV to train the GMM on the training set and obtained the cluster probabilities for every subject in the held-out fold (testing set). Since subjects belonging to the same cluster are likely to share similar clinical profiles, we performed classification by weighing the clinical diagnosis probability per cluster observed in the training data by the probability of belonging to that cluster for each test subject. This yielded an interpretable probability of being assigned each diagnosis. We selected the largest probability as the final classifier prediction.

2.8 | Feature importance

To understand which features are most informative for the classifier when making decisions, we computed feature importance scores. This analysis allowed us to quantify the contribution of each feature to classification performance by nullifying its effect through random permutation of its values across subjects. This effectively removed any meaningful information and associations with other features, creating a null model against which we could compare the classification performance. Each feature was randomly permuted 20 times to arrive at an estimated mean change in classification accuracy. The top 10 features with a mean change in classification accuracy greater than one standard deviation from zero were reported to ensure that reported metrics focused upon results not likely due to chance.

2.9 | Conversion risk

To quantify conversion risk, we first identified the overall rate of conversion to a worse diagnosis at visits 2 and 3, which were within 1.5 years and 3 years of the first visit, respectively. The conversion rate for HC subjects to either MCI or AD dementia was 3.9% at visit 2 and 9.4% at visit 3. The conversion rate for MCI subjects to AD dementia was 27.9% at visit 2 and 39.7% at visit 3. As expected, conversion rates increased over time and were higher for MCI subjects.

To determine whether classifications using visit 1 information revealed information about conversion risk, we stratified HC subjects from visit 1 by classifier prediction. Treating each classifier prediction as a group, we quantified the proportions of subjects within each group that converted to a worse diagnosis at visits 2 and 3. We performed binomial tests to statistically determine whether conversion rates for each prediction group differed from the overall rate for HC subjects at follow-up visits 2 and 3. Binomial tests were only performed for groups that had an expected count of at least five to help ensure reliable results, and *p* values were corrected for multiple comparisons with the Bonferroni correction for each analysis conducted using different feature tiers. We repeated these analyses separately for MCI subjects from visit 1.

We also wanted to understand whether cluster membership revealed patterns of differing conversion risk at visits 2 and 3. In this context, since we did not need to validate predictions from classification, we performed GMM clustering on all subjects using the procedure described earlier. Given that we had five clusters, we performed chi-squared tests to quantify any statistically significant differences in conversion rates across clusters compared to the overall conversion rate for HC subjects from visit 1 at follow-ups 2 and 3. Binomial tests were only performed following significant chi-squared test results with Bonferroni correction to determine which clusters had significantly different conversion rates from the expected overall rate. Binomial test *p* values were also corrected for multiple comparisons with the Bonferroni correction for each analysis conducted using different feature tiers. To ensure appropriate sample sizes, clusters with small counts were combined for both chi-squared tests and post hoc analysis using binomial tests to reach a minimum count of 5. We repeated these analyses separately for MCI subjects from visit 1.

3 | RESULTS

3.1 | Classification performance using cost-effective data

Our first goal was to determine how well the multiclass RF classifier performed when given access to the different tiers of data. Raw evaluation metrics weighted by class prevalence showed a steady increase in performance when higher cost tiers were included in the multiclass RF pipeline (Figure 2A). The overall classification performance when using only tier 1 features achieved 77.3% accuracy (95% confidence interval [CI]: 76.7% to 77.9%) using a RF classification pipeline. When only including non-neuropsychological testing features from tier 1 (ie, excluding the MMSE), the results were highly similar (77.0% accuracy). These values increased to 80.2% accuracy (CI: 79.3% to 81.2%) when the tier 2 set of neuropsychological testing features were introduced. With the inclusion of tier 2 MRI features, the performance metrics increased to 81.3% accuracy (CI: 80.4% to 82.1%). The highest metrics were obtained when tier 3 features were included, yielding 85.5% accuracy (CI: 84.6% to 86.4%). While including all feature tiers produced the most accurate classifications, the relatively modest drops in classification performance when only using lower cost features indicated that using low-cost features for screening is feasible and can yield relatively accurate results that are comparable to neurologists and exceed existing machine learning studies leveraging higher cost data. Balanced evaluation metrics, which average metrics calculated per diagnostic class, were also reported to account for the higher prevalence of cognitively unimpaired individuals in the dataset (Figure 2B). Because MCI diagnoses are often more difficult to correctly classify (Figure 2D) and are less prevalent in the sample, the balanced performance metrics were overall slightly lower than the raw performance metrics that are affected by class size.

3.2 | Comparison of different classifiers

Our next goal was to determine whether our choice of multiclass RF as a baseline approach significantly impacted performance. When using tier 1 features, most of the classification pipelines performed similarly, with our novel hierarchical methods performing significantly better than the multiclass RF. The ensemble method using auto-sklearn performed quite well with an average accuracy of 80.8% (95% CI: 79.8% to 81.8%). This was closely followed by the multiclass RF pipeline as reported earlier (Figure 2C). The three hierarchical pipelines yielded mean accuracy values of 78.8% (CI: 77.7% to 79.9%), 81.5% (CI: 80.0% to 83.0%), and 81.1% (CI: 79.8% to 82.4%) for the dementia-first, unimpaired-first, and combined models, respectively. The latter two models significantly outperformed the multiclass RF model and those in the literature, which used higher cost feature inputs.[5] These models also performed similarly to the auto-sklearn pipeline, which is an ensemble method and much more computationally expensive to train. This demonstrates the potential of our hierarchical strategy to effectively leverage lower-cost data. The clustering-classification pipeline yielded a mean accuracy of 76.0% (CI: 75.3% to 76.8%) and was comparable to the RF pipeline. These results suggest that the reported performance metrics using tier 1 features are reliable across a variety of classification strategies.

We further analyzed several binary classification tasks for classifying HC, MCI, and AD dementia using tier 1 features (Figure 2D). Using the RF pipeline, we achieved accuracies of 96.2% for AD dementia versus HC, 91.9% for AD dementia versus other (screening for AD dementia), and 88.9% for HC versus other (screening for cognitive impairment due to AD), as shown in Figure 2D. In the latter two tasks, we achieved an AD dementia screening F1-score of 0.919 (95% CI: 0.916 to 0.922) and a cognitive impairment screening F1-score of 0.889 (95% CI: 0.884 to 0.894). Classification accuracies were similar for the auto-sklearn pipeline. These results are comparable to those in published studies that leveraged more costly features.[43] The highest accuracies were associated with distinguishing AD dementia, HC, or both, while the more difficult tasks involved MCI classification.

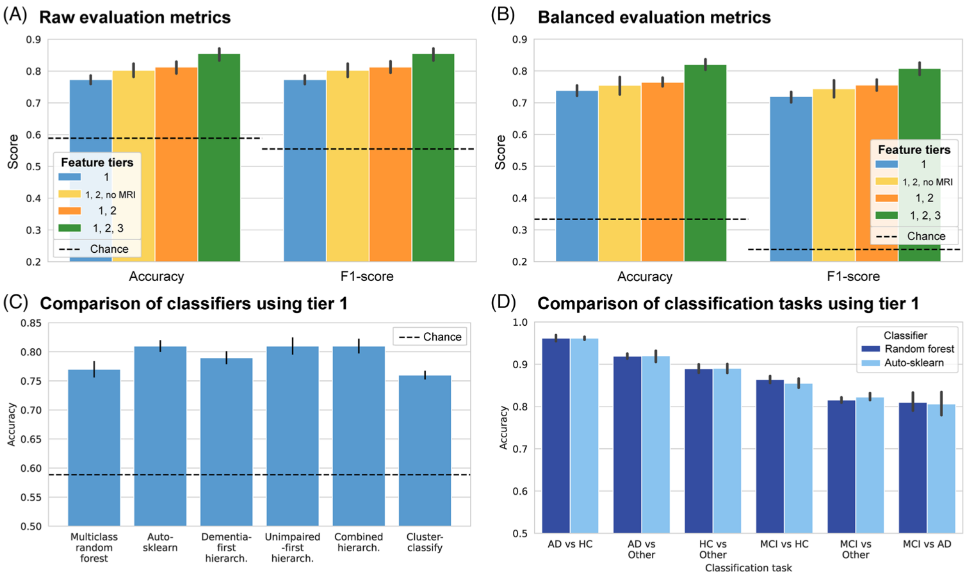

FIGURE 2 Comparison of classification performance using different feature tiers and classification pipelines. (A) Raw metrics increased with added feature tiers. Chance line indicates the performance of a naïve classifier that always predicts the most frequent class label. (B) Balanced metrics followed the same trend, with slightly lower scores due to emphasis on the MCI class. (C) Performance across all pipelines, including hierarchical and clustering-classification models, was similar. Hierarchical strategies outperformed the multiclass RF pipeline. (D) The highest binary classification accuracy was associated with distinguishing AD dementia and HC individuals. Error bars represent 95% CI across five-fold CV.

3.3 | Identifying important features in AD screening

To determine which data features might be most clinically relevant, we investigated the feature importance values associated with classifying clinical diagnoses using different tiers of features. When only tier 1 features were used in the RF classification pipeline (Figure 3A), the top three features ranked in terms of contribution to classification accuracy were each person's subjective report of memory decline (DECSUB), the MMSE score, and whether the subject had any difficulty regarding travel and driving (TRAVEL). Many of the top ranked features were from the patient interview and reflected a combination of functional ability (FAS features: TRAVEL, REMDATES, TAXES), living situation (level of independence and marital status), and demographic information.

When tier 2 neuropsychological testing features were included, the top features were a mix between tier 1 and tier 2 neuropsychological testing features (Figure 3B). The subjects' delayed recall of story units (MEMUNITS) and performance in two categorical fluency tests (VEG and ANIMALS) were among the top four features. DECSUB remained a top feature, along with functional ability (TRAVEL and BILLS). MMSE was ranked seventh, followed by more tier 1 features.

Including tier 2 MRI features revealed a similar trend of having MEMUNITS and DECSUB in the top two spots of feature importance (Figure 3C). The same tier 2 neuropsychological testing features (delayed recall, vegetable and animal naming) were among the top 10. The remaining top 10 features were from tier 1. Cerebrum CSF volume (CERECSF), left supramarginal gray matter volume, and left paracentral gray matter volume were among the top 20 features ranked by importance.

When all three tiers of features were included, tier 3 features were among the most important features (Figure 3D). The global CDR score (CDRGLOB) and memory domain subscore (MEMORY) were ranked particularly high, followed by MEMUNITS from tier 2 and DECSUB from tier 1. Given that the CDR scores are highly correlated with the clinical diagnosis, this ranking of features was unsurprising. CDR sum of boxes (CDRSUM) came in at fifth, likely due to the presence of all the CDR domain subscores that carried the same information. CERECSF was among the top 10 and closely followed by the left parahippocampal mean cortical thickness and left middle temporal mean cortical thickness.

To understand the information carried by each feature in relation to classification performance, we conducted classification using each of the top 10 informative features by itself (Figure S1). For the top tier 1 features, several yielded raw accuracies that were not far behind our base model, as the diagnostic class imbalance of our population could drive accuracy, but all trailed the base model considerably when this was accounted for in the balanced accuracy measure. This remained similar for the top tier 2 features. For the top tier 3 features, we observed higher accuracy and balanced accuracy results. Across all features, the balanced accuracy was consistently much lower than the accuracy, suggesting that single feature predictors struggled to accurately classify MCI subjects and were likely better at capturing HC (and perhaps AD) subjects, as MCI-related classifications tend to be more difficult (Figure 2D). The classification performance of each of the top tier 3 features is similar to that of all tier 1 features combined, which helps to explain the modest improvement in

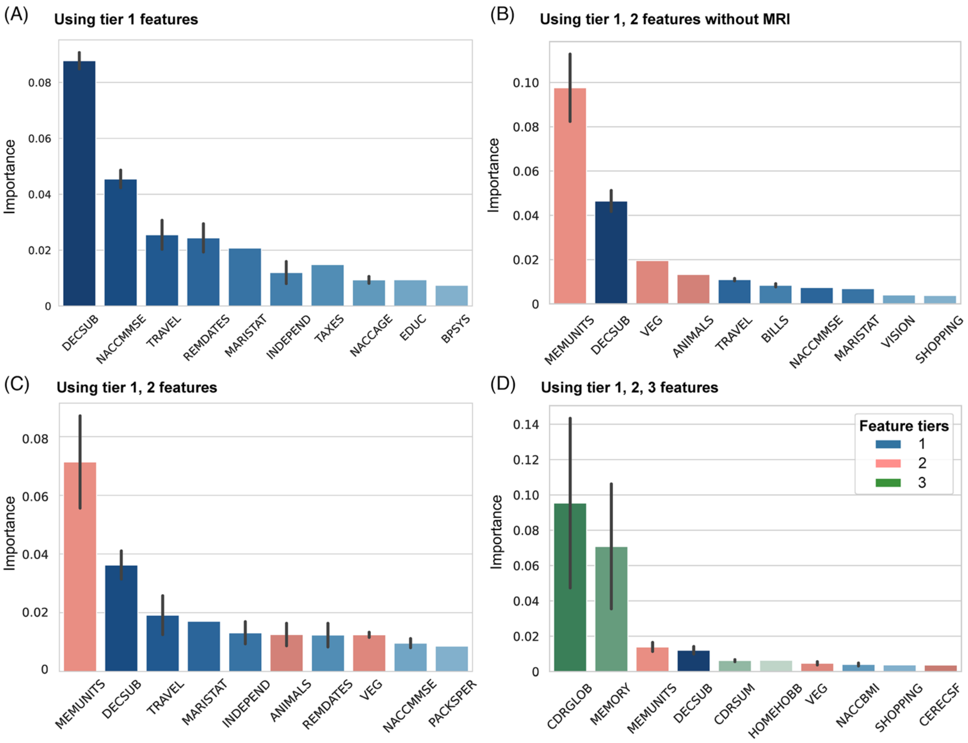

FIGURE 3 Identifying important features when using tier 1 only (A), tiers 1 and 2 without MRI (B), tiers 1 and 2 (C), and tiers 1 to 3 (D) features. Features are colored as follows: blue: tier 1, pink: tier 2, and green: tier 3, and error bars represent SEM across CV folds. (A) Tier 1-only classification was most impacted by the subjects' self-report of significant memory decline (DECSUB). (B) With added tier 2 neuropsychological testing features, memory again was highly relevant (delayed recall of story units, DECSUB from tier 1, and performance in two categorical fluency tests). (C) Including MRI revealed a similar trend. MRI features did not rank among the top 10, but cerebrum CSF volume (CERECSF), left supramarginal gray matter volume, and left paracentral gray matter volume were among the top 20 features. (D) When all three tiers were included, tier 3 features were among the most important features. Detailed descriptions of feature definitions can be found under Figure S1.

classification performance when including tier 3 features in addition to tier 1 and 2 features.

3.4 | Baseline classifications significantly inform conversion rate

While misclassifications are commonly treated as mistakes, we know that clinical diagnoses of AD are imperfect, and it is certainly possible that clinical diagnoses lag the true patient status. We hypothesized that classifications, especially misclassifications of baseline diagnoses, may provide insight into follow-up diagnoses and, subsequently, conversion risk of individuals at later time points. To determine whether this was the case, we used the classifier predictions from the first visit and compared the results to follow-up diagnoses received by subjects at visit 2 (within 1.5 years) and visit 3 (within 3 years) of the first visit.

In Figure 4, correctly classified subjects are in darker colors and incorrectly classified subjects are in lighter colors, and baseline conversion rates are indicated by a dashed line. When only tier 1 features were used, conversion rates for correctly predicted subjects diagnosed as HC at visit 1 were not statistically different from the sample conversion rate at visit 2 or 3 (Figure 4A, left). Incorrectly predicted subjects had conversion rates that trended higher, but, due to limited sample sizes, no statistical tests were performed. For subjects diagnosed with MCI at visit 1, 27.9% converted to AD dementia at visit 2 and 39.7% converted to AD dementia at visit 3 (Figure 4A, right). In contrast to these rates, individuals who were mistakenly classified as HC at visit 1 had lower conversion rates, and individuals mistakenly classified with AD dementia had significantly higher conversion rates (light blue, $p < .05$). Correctly classified individuals' conversion rates did not differ from the base rate (dark blue). This effect was most prominently observed in visit 2 due to sample size limitations for visit 3 subjects.

In a similar fashion, we used tier 1 and 2 features without MRI to classify visit 1 diagnoses and quantified HC subjects' conversion rates by classifier prediction (Figure 4B, left). Correctly classified subjects mimicked the sample conversion rate, and misclassified individuals trended toward higher conversion rates to MCI or AD dementia at follow-up. Similarly, for MCI subjects at visit 1, correctly classified individuals represented the sample rate of conversion, while misclassified subjects had lower or significantly higher rates of conversion at visit 2 (Figure 4B, right, light orange, $p < .05$). We observed similar trends when incorporating MRI features and tier 3 features (Figure S2). Taken together, these results indicate that misclassifications are, in fact, predictive of both positive and negative changes in subsequent conversion rates.

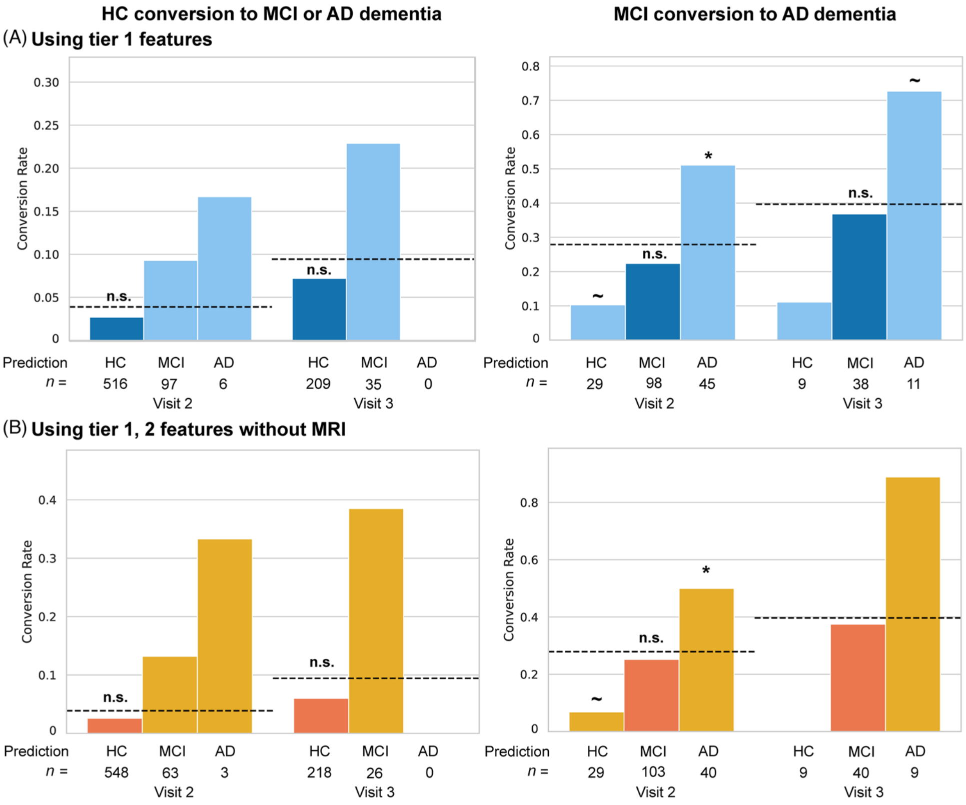

FIGURE 4 Predictions of visit 1 diagnoses reveal significant insight into conversion rates at follow-up visits. (A) Left: Using tier 1 features only, predictions for HC subjects trended toward higher conversion rates for misclassified subjects (light blue) compared to the sample rate (dashed lines). (A) Right: For subjects diagnosed as MCI at visit 1, individuals mistakenly predicted as HC had lower conversion rates and individuals mistakenly predicted to have AD dementia had higher conversion rates, especially in visit 2 (light blue, $p < .05$). (B) Left: Using tier 1 and 2 features without MRI, correctly predicted HC subjects (dark orange) followed the sample rate, while misclassified individuals trended toward higher conversion rates at follow-up (light orange). (B) Right: For MCI subjects, correctly predicted individuals represented the sample conversion rate (dark orange), while misclassified subjects had different rates of conversion, significant for subjects incorrectly classified as AD dementia (light orange, $p < .05$). For binomial tests, * denotes $p < .05$ after Bonferroni correction; ~ denotes $p < .05$ prior to correction; n.s. denotes non-significant; no notation indicates that no test was conducted due to small sample size.

3.5 | Clustering identified high-risk groups with elevated conversion rates

While classification performance using the clustering-classification pipeline was comparable to the baseline RF classifier and lower than the hierarchical models, this method does provide additional information in the form of the cluster membership of each subject. We hypothesized that the cluster memberships might reveal different clinical subtypes within the AD spectrum that show differential rates of conversion at follow-up visits. To determine whether lower cost features perform similarly to more costly features, we clustered subjects using tier 1 features (Figure 5A), tier 1 and 2 features without MRI data (Figure 5B), tier 1 and 2 features including MRI (Figure S3C), or tier 1 through 3 features (Figure S3D). We hypothesized that the cluster memberships using cost-effective features would reveal similar trends as when using higher-cost features.

Using only tier 1 features produced five clusters, each of which showcased meaningful diagnosis compositions (Figure 5A, left). Cluster A contained mostly cognitively unimpaired individuals, while clusters D and E were composed mostly of AD dementia subjects. Clusters B and C were more heterogeneous across diagnosis groups. For subjects diagnosed as HC from visit 1, conversion rates differed significantly from the sample rate for subjects in clusters C to E at visit 2, who had higher rates of conversion at 13.3% compared to the baseline rate of 3.9% (Figure 5A, middle; $p < .05$). At visit 3, this group of HC subjects continued to exhibit significantly high conversion rates of 75.0% compared to the sample rate of 9.4% (Figure 5A, right; $p < .05$). This subset of HC subjects within clusters C to E is at significantly elevated risk—more than three times more likely to decline at visit 2 and almost eight times more likely to decline at visit 3 than the population—and should be prioritized for follow-up care in the clinic.

For MCI subjects, conversion rates were significantly higher for subjects in cluster D at visit 2 (66.7% compared to 27.9%, $p < .005$). For subjects in cluster A, conversion rates were lower than the sample rate at visit 2 (10.8% compared to 27.9%, marginally significant). These results showed that MCI subjects belonging in cluster A were less likely to progress to AD dementia at follow-up, but subjects in cluster D were at significantly higher risk of progressing to AD dementia (Figure 5A, right). This group of high-risk MCI subjects is more than two times

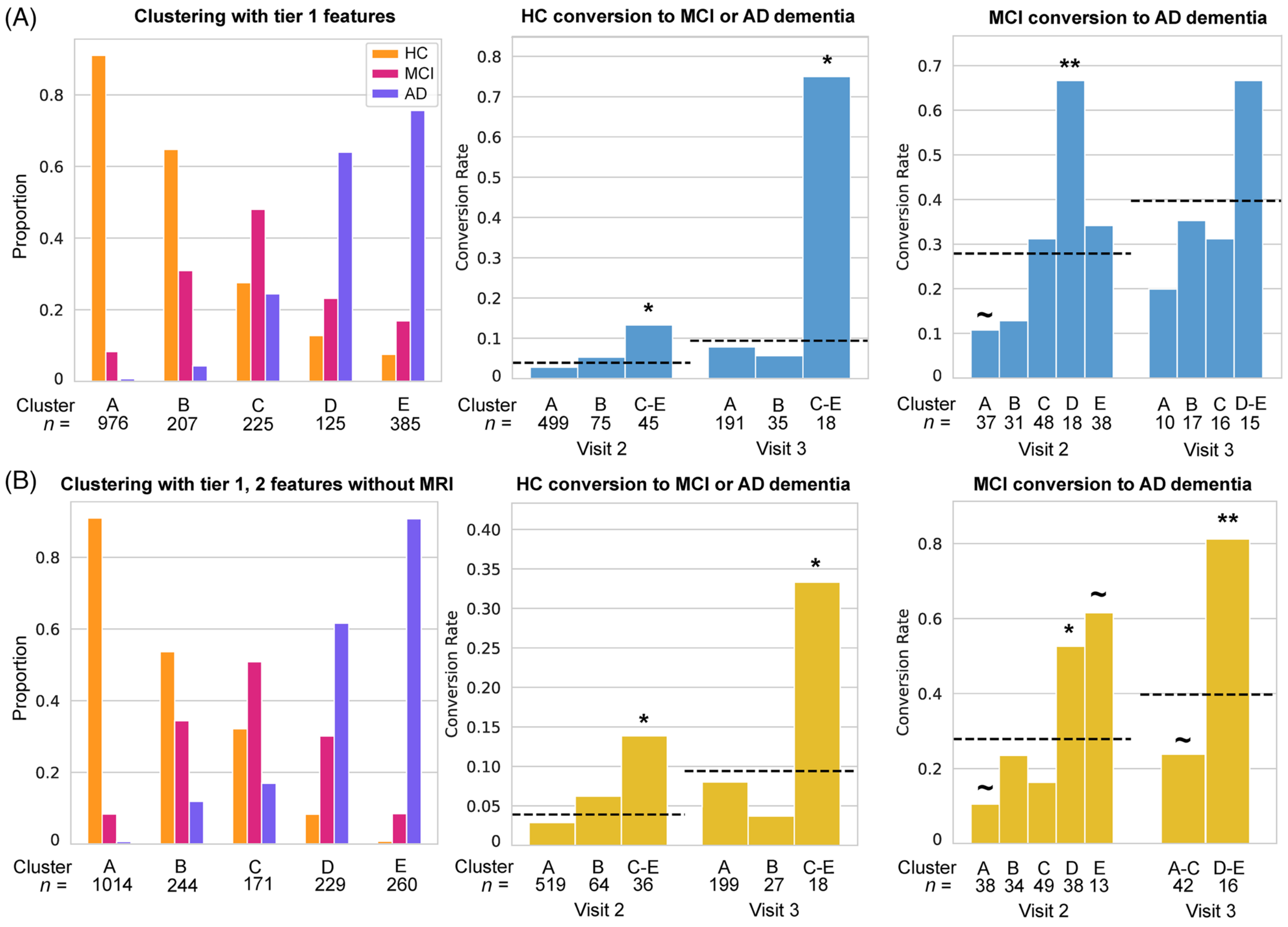

FIGURE 5 Cluster membership is linked to risk of subsequent decline. (A) Using tier 1 features only, for HC subjects, cluster B membership indicated significantly greater conversion at visit 2 compared to expected rate shown as the dashed line ($p < .05$), and cluster C to E membership indicated significantly greater conversion at visit 3 compared to expected ($p < .05$). For MCI subjects, cluster A was associated with lower conversion (marginally significant), while cluster C was associated with significantly higher conversion to AD dementia at visit 2 ($p < .005$). (B) Using tier 1 and 2 features without MRI, for HC subjects, cluster C to E membership was associated with significantly greater conversion at both visits 2 and 3 compared to the expected rates ($p < .05$). For MCI subjects, cluster D membership was associated with significantly greater conversion at visit 2 ($p < .05$) and cluster C to E membership was associated with significantly greater conversion at visit 3 ($p < .005$). For binomial tests, * indicates $p < .05$ and ** denotes $p < .005$ after Bonferroni correction; ∼ denotes $p < .05$ prior to correction.

more likely to decline at visit 2 compared to their peers and should be prioritized for clinical care.

Using tier 1 and 2 features without MRI, we produced five clusters with similar distributions of clinical diagnoses (Figure 5B, left). Clusters A and B were mostly composed of cognitively unimpaired individuals. Cluster C contained mostly MCI subjects, and clusters D and E were predominantly represented by AD dementia individuals. HC subjects belonging to clusters C to E had significantly greater conversion rates at follow-up (13.9% vs baseline 3.9% at visit 2 with $p < .05$, 33.3% vs baseline 9.4% at visit 3 with $p < .05$). This shows that cognitively unimpaired individuals in clusters C to E are almost four times more likely to worsen at follow-up visits (Figure 5B, middle). For MCI subjects, those in cluster D had significantly elevated conversion rates at visit 2 (52.6% vs. 27.9%, $p < .05$), and those in clusters D or E had significantly elevated conversion rates at visit 3 compared to the sample rate (81.3% vs. 39.7%, $p < .005$). This represents a nearly two-fold increase in conversion rate at visits 2 and 3 compared to the population rate (Figure 5B, right). These individuals should be prioritized for follow-up clinical visits to ensure proper care is delivered. We reported similar results when incorporating MRI and tier 3 features (Figure S3).

4 | DISCUSSION

In this study, we investigated the application of statistical machine learning to understanding AD clinical diagnoses using cost-effective data from NACC. We achieved high accuracy (∼90%) for the most vital binary classifications (HC vs. other and AD vs. other) and moderately high accuracy (∼80%) for identifying MCI using low-cost tier 1 features. This indicates that using lower cost data holds promise for identifying even the earliest clinical symptoms. We further demonstrated the utility of low-cost data by achieving multinomial classification results that were more accurate than those of past studies (which used higher cost features), and we identified the most important features contributing to this. Notably, the RF classifier's labels were more indicative of conversion risk than would be expected by chance, and our clustering method identified high-risk subgroups that had significantly greater conversion risk at follow-up visits. This work establishes a framework for more efficient AD screening in tandem with quantification of conversion risk, conferring confidence in classification using cost-effective data.

Our reported classification performance is very competitive with existing work. Of the studies that utilized multimodal data, some did report classification accuracy using a subset of modalities. El-Sappagh et al. reported a RF multiclass classification accuracy of 77.8% using neuropsychological testing features and, when using medical history features, an accuracy of 46.2%.[4] In contrast, when we utilized tier 1 features (without neuropsychological tests beyond the simple MMSE), we obtained a multiclass accuracy of 77.3% using RF and 81.5% using our combined hierarchical pipeline. Qiu et al. reported deep learning multiclass accuracy of 78.2% using medical history, clinical data, functional assessment, and neuropsychological testing features.[5] In comparison, our results using tier 1 and 2 features (without MRI) yielded a multiclass accuracy of 80.2% using a simpler RF classifier.

Intuitively, one might presuppose that by including tier 3 features, accuracy could reach 100% as this includes the full set of CDR scores often used to define disease status in the literature.[9,44] However, the clinical diagnosis is more complex and considers a wider range of factors. In our sample, global CDR scores and CDR sum of boxes scores overlapped considerably across the three clinical diagnoses and, on their own, only obtained balanced accuracies of 71% to 72% (Figure S1D), notably lower than the 82.1% achieved by the RF classifier when using tier 1 to 3 features. Additionally, clinical experts can disagree on diagnoses, quantified by meta-analyses on diagnostic criteria used in the NACC UDS,[22,23] which showed that agreement for diagnoses of AD-related cognitive impairment yielded a kappa of 0.71 (95% CI: 0.65 to 0.77).[45] When specifically using the NIH-AA 2011 diagnostic criteria, a study found an average kappa of 0.76 (95% CI: 0.65 to 0.86) between neurologists.[46] Our RF model yielded a similar kappa score of 0.75 when using all feature tiers. Given that our model mimics experts when producing clinical diagnoses, the accuracy we achieved using tier 3 features may represent a near-ceiling level of performance. In this context, the classification performance using tier 1 features is impressive and suggest that tier 1 features, when used aptly, can capture a large amount of the signal carried by the more costly tier 3 features.

Our binary classification results exceeded the screening accuracy of general practitioners. Using only tier 1 features, we produced very high accuracies that were comparable to past work, which used higher cost data not readily available to most older adults.[47] In the context of primary care, meta-analyses found that general practitioners achieved an F1-score of 0.735 when screening for dementia and an F1-score of 0.785 when screening for cognitive impairment.[2] Our pipeline, using only primary care-accessible features (tier 1), outperformed these findings, with an AD dementia screening F1-score of 0.919 and a cognitive impairment screening F1-score of 0.889. These results were robust across classification methodology. Beyond the scope of general medical practice, our model was also able to capture MCI status at a high level of accuracy using tier 1 features (81.5% for MCI vs. other, 86.4% for MCI vs. HC, and 81.0% for MCI vs. AD dementia). Thus, cost-effective data can produce very accurate screening results for AD dementia and AD-related cognitive impairment even at the earliest stages.

In addition, we developed novel hierarchical models that performed competitively with the more complex ensemble pipeline. Given that the MCI classification task was the hardest, we hypothesized that we could improve classification performance by breaking down the multiclass landscape into two sequential binary classifications. From this we developed the dementia-first, unimpaired-first, and combined hierarchical models. When using tier 1 features, the latter two methods, which only used two classifiers each, performed similarly to the ensemble auto-sklearn pipeline, which used up to seven classifiers and was much more computationally expensive. This shows that applying relevant domain knowledge can significantly aid diagnosis prediction.

An alternative approach to classification is clustering of data in an unsupervised manner. Classification using our clustering-based pipeline achieved an accuracy of 75.8%, outperforming past work that used k-nearest neighbors and achieved a multiclass accuracy of 64.76% using a larger number of modalities, including neuropsychological testing, CDR scores, PET, MRI, and CSF biomarkers.[4] We also gained further insight into conversion risk associated with cluster membership, which can guide triaging of patients at higher or lower risk of decline over time. We expect that future work with larger sample sizes for longitudinal visits will lead to more significant findings and replicate our results, which were limited by sample size.

Using the foregoing classification results, we developed novel analyses to gain insight into conversion risk at follow-up visits using only baseline visit information. First, we found that the classifier's predictions, especially for misclassified subjects, were significantly indicative of conversion than the expected population rate. This was observed for both the RF and hierarchical classifiers. This suggests that the classifiers are leading the clinical diagnosis and are sensitive to features predictive of conversion risk. Second, our clustering model represents a promising approach for finding clinically meaningful subgroups while also providing classification results. With only tier 1 features, we identified a group of high-risk HC subjects who were three times more likely to decline at visit 2 and almost eight times more likely to decline at visit 3. We identified another subset of high-risk MCI subjects who were more than two times more likely to decline at visit 2 compared to their peers. In future work, we aim to further analyze these results to identify the clinical profiles of subjects within each cluster. This will help determine whether there are differential features across clusters that may help us understand cluster membership.

To better understand our findings, we investigated which features were important to the classification. The subject's self-report of noticeable memory decline (DECSUB) was consistently one of the top informative features. Further exploration is required to tease apart the relationships between DECSUB and the often-present neuropsychiatric variables in this population[48] and to understand their relationship to the classification results. In light of our findings that the classifier's predictions led clinical diagnosis, DECSUB and other important features may be predictive features of subsequent decline. In contrast, the included MRI features were not important to our classifiers. Including them in tier 2 did not improve performance, and none were ranked within the top 10 features. We note that these are basic structural features from NACC and do not necessarily represent structural MRI's true potential or the potential for other forms of MRI (eg, ASL, DWI, MRS, rsfMRI). Additionally, scanner and acquisition differences across ADRCs may reduce the signal extracted from MRI here.

We further examined feature importance by evaluating classification accuracy using individual features (Figure S1). We observed that the drop in balanced accuracy compared to raw accuracy was much greater than expected from earlier results (Figure 2). This suggests that single feature predictors struggled to accurately classify the more challenging MCI subjects (Figure 2D). This drop in balanced accuracy was most noticeable for tier 1 and 2 features. Classification using the CDR global score (CDRGLOB) achieved raw (78.0%) and balanced (73.2%) accuracies comparable to those obtained using tier 1 features. This was also observed when using the CDR sum of boxes score (raw: 82.0%, balanced: 71.7%). These CDR scores are used as key indicators of outcome in research studies and clinical trials for AD and related dementias.[44] Additionally, prior work showed that machine learning classifiers could accurately predict categories of CDR scores using tier 1 and 2 features.[9] Yet, the combination of tier 1 features was able to perform on par with these gold standards, offering a far more accessible and cost-effective solution.

Several notable limitations exist in our work. Since we based our classification targets on clinical diagnosis from experts, classification metrics are not reflective of *post mortem* neuropathological diagnosis of AD. While we defined tier 1 as data that could be reasonably collected from a primary care visit, we recognize that this may be constrained by different resource and time limitations to only include a subset of tier 1 features. Future work to investigate how well limited sets of tier 1 features can perform will help elucidate this. As this study focused on late-onset AD, this work limited its scope to older adults above the age of 65, which can be a limitation when considering earlier onset ages for AD. While we had a large cross-sectional sample for the first visit, longitudinal sample sizes were limited in this study. This limited the number of statistical tests we could conduct to evaluate the significance of conversion rate findings. Additionally, the NACC dataset is not a representative sample of the aging population in the United States due to selection and volunteer biases. More work must be done in future studies to address the lack of representation of different groups and to improve diversity, equity, and inclusion in aging studies. We have identified other datasets with lower cost features to further validate these findings in more representative samples that have greater ethnic, racial, and geographic representation available from Medicare's Annual Wellness Visit or from the Health and Retirement Study and related studies.[49–51] With increased access to plasma biomarkers and the recent draft of the "NIA-AA Revised Criteria for Diagnosis and Staging of Alzhiemer's Disease," we envision that plasma biomarkers, like other tier 1 features, can complement the existing screening algorithms to allow for predictions of both clinical and biological AD staging. Our future work includes incorporating plasma biomarkers to improve the diagnostic utility of lower cost data for more older adults.

ACKNOWLEDGMENTS

We thank the University of California, Irvine Alzheimer's Disease Research Center's experts for their insights into the feature cost tiers. We are grateful to the National Alzheimer's Coordinating Center (NACC) for generously sharing various forms of clinical and imaging data to empower research. The NACC database is funded by National Institute on Aging (NIA)/National Institutes of Health (NIH) Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), and P30 AG072959 (PI James Leverenz, MD). The authors acknowledge funding support from the National Institute on Aging (Grants P30AG066519 and F30AG079610) and National Institute of Mental Health (Grant R01MH115697) of the NIH.

CONFLICT OF INTEREST STATEMENT

The authors report no conflicts of interest. Author disclosures are available in the supporting information.

CONSENT STATEMENT

All human subject data used in this study were obtained from NACC and recorded such that subjects cannot be identified. For this reason, consent was not necessary for the use of human subject data in this study.

ORCID

Yueqi Ren https://orcid.org/0000-0003-2936-6009

REFERENCES

1. Alzheimer's Association. 2023 Alzheimer's Disease facts and figures. *Alzheimers Dement*. 2023;19:1598-1695. doi:10.1002/alz.13016
2. Creavin ST, Noel-Storr AH, Langdon R, et al. Clinical judgement by primary care physicians for the diagnosis of all-cause dementia or cognitive impairment in symptomatic people. Cochrane Dementia and Cognitive Improvement Group, ed. *Cochrane Database Syst Rev*. 2022;2022(6). doi:10.1002/14651858.CD012558.pub2
3. Suk HI, Lee SW, Shen D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage*. 2014;101:569-582. doi:10.1016/J.NEUROIMAGE.2014.06.077
4. El-Sappagh S, Alonso JM, Islam SMR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep*. 2021;11(1):2660. doi:10.1038/s41598-021-82098-3
5. Qiu S, Miller MI, Joshi PS, et al. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat Commun*. 2022;13(1):3404. doi:10.1038/s41467-022-31037-5
6. Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*. 2014;8:14. doi:10.3389/fninf.2014.00014
7. Bron EE, Smits M, van der Flier WM, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *Neuroimage*. 2015;111:562-579. doi:10.1016/j.neuroimage.2015.01.048
8. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans Pattern Anal Mach Intell*. 2020;42(4):880-893. doi:10.1109/TPAMI.2018.2889096
9. for the Alzheimer's Disease Neuroimaging Initiative, Kleiman K, Barenholtz E, Galvin JE. Screening for early-stage Alzheimer's disease using optimized feature sets and machine learning. *JAD*. 2021;81(1):355-366. doi:10.3233/JAD-201377
10. Weakley A, Williams JA, Schmitter-Edgecombe M, Cook DJ. Neuropsychological test selection for cognitive impairment classification: a machine learning approach. *J Clin Exp Neuropsychol*. 2015;37(9):899-916. doi:10.1080/13803395.2015.1067290
11. Bogdanovic B, Eftimov T, Simjanoska M. In-depth insights into Alzheimer's disease by using explainable machine learning approach. *Sci Rep*. 2022;12(1):6508. doi:10.1038/s41598-022-10202-2
12. Pereira T, Ferreira FL, Cardoso S, et al. Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease: a feature selection ensemble combining stability and predictability. *BMC Med Inform Decis Mak*. 2018;18(1):137. doi:10.1186/s12911-018-0710-y
13. Na KS. Prediction of future cognitive impairment among the community elderly: a machine-learning based approach. *Sci Rep*. 2019;9(1):3335. doi:10.1038/s41598-019-39478-7
14. Cleret DE, Langavant L, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. *J Med Internet Res*. 2018;20(7):e10493. doi:10.2196/10493
15. Schlesinger DE, Stultz CM. Deep learning for cardiovascular risk stratification. *Curr Treat Options Cardio Med*. 2020;22(8):15. doi:10.1007/s11936-020-00814-0
16. Albright J, Alzheimer's Disease Neuroimaging Initiative. Forecasting the progression of Alzheimer's disease using neural networks and a novel preprocessing algorithm. *Alzheimer's Dement*. 2019;5(1):483-491. doi:10.1016/j.trci.2019.07.001
17. Mofrad SA, Lundervold A, Lundervold AS. A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Comput Med Imaging Graph*. 2021;90:101910. doi:10.1016/j.compmedimag.2021.101910
18. Casanova R, Saldana S, Lutz MW, Plassman BL, Kuchibhatla M, Hayden KM. Investigating predictors of cognitive decline using machine learning. *J Gerontol B Psychol Sci Soc Sci*. 2020;75(4):733-742. doi:10.1093/geronb/gby054
19. Grassi M, Rouleaux N, Caldirola D, et al. 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;10(JUL):756. doi:10.3389/fneur.2019.00756
20. Katabathula S, Davis PB, Xu R. Comorbidity-driven multi-modal subtype analysis in mild cognitive impairment of Alzheimer's disease. *Alzheimer's Dement*. 2023;19(4):1428-1439. doi:10.1002/alz.12792
21. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the uniform data set. *Alzheimer Dis Assoc Disord*. 2007;21(3):249-258. doi:10.1097/WAD.0b013e318142774e
22. Morris JC, Weintraub S, Chui HC, et al. The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Dis Assoc Disord*. 2006;20(4):210-216. doi:10.1097/01.wad.0000213865.09806.92
23. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's Coordinating Center's uniform data set. *Alzheimer Dis Assoc Disord*. 2018;32(4):351-358. doi:10.1097/WAD.0000000000000279
24. Weintraub S, Besser L, Dodge HH, et al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Dis Assoc Disord*. 2018;32(1):10-17. doi:10.1097/WAD.0000000000000223
25. Monsell SE, Dodge HH, Zhou XH, et al. Results from the NACC uniform data set neuropsychological battery crosswalk study. *Alzheimer Dis Assoc Disord*. 2016;30(2):134-139. doi:10.1097/WAD.0000000000000111
26. DeCarli C. Updated MRI methods. Accessed May 3, 2023. https://files.alz.washington.edu/documentation/updated-mri-methods.pdf
27. Creavin ST, Wisniewski S, Noel-Storr AH, et al. Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. Cochrane Dementia and Cognitive Improvement Group, ed. *Cochrane Database Syst Rev*. 2016(1). doi:10.1002/14651858.CD011145.pub2
28. Cordell CB, Borson S, Boustani M, et al. Alzheimer's Association recommendations for operationalizing the detection of cognitive impairment during the Medicare Annual Wellness Visit in a primary care setting. *Alzheimers Dement*. 2013;9(2):141-150. doi:10.1016/j.jalz.2012.09.011
29. Mayeux R, Saunders AM, Shea S, et al. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. *N Engl J Med*. 1998;338(8):506-511. doi:10.1056/NEJM199802193380804
30. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953
31. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24(6):417-441. doi:10.1037/h0071325
32. Tibshirani R. Regression shrinkage and selection via the lasso. *JSTOR*. 1996;58(1):267-288.
33. Tipping ME. Sparse Bayesian Learning and the relevance vector machine. *JMLR*. 2001;1:211-244.
34. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:10.1007/s10994-006-6226-1
35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.
36. Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2016;18(17):1-5.
37. Breiman L. Random forests. *Machine Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

38. Lebedev AV, Westman E, Van Westen GJP, et al. Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *Neuroimage Clin*. 2014;6:115-125. doi:10.1016/j.nicl.2014.08.023
39. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In: *Advances in Neural Information Processing Systems*. Vol 28. Curran Associates, Inc.; 2015.
40. Gamberger D, Lavrač N, Srivatsa S, Tanzi RE, Doraiswamy PM. Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. *Sci Rep*. 2017;7(1):6763. doi:10.1038/s41598-017-06624-y
41. Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The application of unsupervised clustering methods to Alzheimer's disease. *Front Comput Neurosci*. 2019;13:31. doi:10.3389/fncom.2019.00031
42. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7
43. Ezzati A, Zammit AR, Harvey DJ, Habeck C, Hall CB, Lipton RB. Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *J Alzheimers Dis*. 2019;71(3):1027-1036. doi:10.3233/JAD-190262
44. Cedarbaum JM, Jaros M, Hernandez C, et al. Rationale for use of the clinical dementia rating sum of boxes as a primary outcome measure for Alzheimer's disease clinical trials. *Alzheimers Dement*. 2013;9(1S):S45-S55. doi:10.1016/j.jalz.2011.11.002
45. Cerullo E, Quinn TJ, McCleery J, Vounzoulaki E, Cooper NJ, Sutton AJ. Interrater agreement in dementia diagnosis: a systematic review and meta-analysis. *Int J Geriatr Psychiatry*. 2021;36(8):1127-1147. doi:10.1002/gps.5499
46. Llamas-Velasco S, Sierra-Hidalgo F, Llorente-Ayuso L, Herrero-San Martín A, Villarejo Galende A, Bermejo Pareja F. Inter-rater agreement in the clinical diagnosis of cognitive status: data from the neurological disorders in Central Spain 2 pilot study. *Neuroepidemiology*. 2016;47(1):32-37. doi:10.1159/000447699
47. Ezzati A, Zammit AR, Harvey DJ, Habeck C, Hall CB, Lipton RB. Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *JAD*. 2019;71(3):1027-1036. doi:10.3233/JAD-190262
48. Lyketsos CG, Lopez O, Jones B, Fitzpatrick AL, Breitner J, DeKosky S. Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. *JAMA*. 2002;288(12):1475. doi:10.1001/jama.288.12.1475
49. Gorin SS, Resnick B. Introduction to the annual wellness visit for the older adult. *PPAR*. 2019;29(1):1-4. doi:10.1093/ppar/pry052
50. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort profile: the health and retirement study (HRS). *Int J Epidemiol*. 2014;43(2):576-585. doi:10.1093/ije/dyu067
51. Langa KM, Ryan LH, McCammon RJ, et al. The health and retirement study harmonized cognitive assessment protocol project: study design and methods. *Neuroepidemiology*. 2020;54(1):64-74. doi:10.1159/000503004

SUPPORTING INFORMATION

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

How to cite this article: Ren Y, Shahbaba B, Stark CEL. Improving clinical efficiency in screening for cognitive impairment due to Alzheimer's. *Alzheimer's Dement*. 2023;15:e12494. https://doi.org/10.1002/dad2.12494