Efficient and Explainable Risk Assessments for Imminent Dementia in an Aging Cohort Study

Nicasia Beebe-Wang*, Alex Okeson*, Tim Althoff**, and Su-In Lee**

***Abstract*— As the aging US population grows, scalable approaches are needed to identify individuals at risk for dementia. Common prediction tools have limited predictive value, involve expensive neuroimaging, or require extensive and repeated cognitive testing. None of these approaches scale to the sizable aging population who do not receive routine clinical assessments. Our study seeks a tractable and widely administrable set of metrics that can accurately predict imminent (i.e., within three years) dementia onset. To this end, we develop and apply a machine learning (ML) model to an aging cohort study with an extensive set of longitudinal clinical variables to highlight at-risk individuals with better accuracy than standard rudimentary approaches. Next, we reduce the burden needed to achieve accurate risk assessments for those deemed at risk by (1) predicting when consecutive clinical visits may be unnecessary, and (2) selecting a subset of highly predictive cognitive tests. Finally, we demonstrate that our method successfully provides individualized prediction explanations that retain non-linear feature effects present in the data. Our final model, which uses only four cognitive tests (less than 20 minutes to administer) collected in a single visit, affords predictive performance comparable to a standard 100-minute neuropsychological battery and personalized risk explanations. Our approach shows the potential for an efficient tool for screening and explaining dementia risk in the general aging population.**

***Index Terms*— dementia, feature selection, geriatrics, interpretability, personalized medicine**

I. Introduction

ALZHEIMER'S disease (AD), a degenerative brain condition, affects an estimated 5.8 million Americans. As the world's older population grows at an unprecedented rate, the number of individuals with dementia is projected to more than double, making it an increasingly pressing health concern [1]. Significant advances in diagnostic predictions are essential to curb the devastating effects of dementia worldwide. We believe these advances will be enabled by large-scale aging cohort studies and machine learning (ML) innovations.

Although no currently known treatment can cure or retard AD progression, identifying AD cases before severe neurological damage ensues is crucial. Predicting onset can promote treatment efficacy once successful interventions are developed and swiftly identify individuals who may benefit from drug trials. It will also help families plan for patient care and patients to receive resources to help make personal decisions about their care before they lose the autonomy to do so [2].

Although studies have demonstrated the possibility of identifying individuals who already have dementia [3], such diagnoses occur beyond the critical window for effective interventions or end-of-life planning [2]. Other studies have predicted the onset of dementia in advance of a clinical diagnosis, but often involve costly data collection using neuroimaging or in-depth neuropsychological batteries over multiple years [4]–[9]. The use of repeated cognitive testing may help to model and predict an individual's cognitive decline [8]; however, given that only 16% of American seniors receive regular cognitive assessments in primary care settings [10], this approach may be impractical for the general population. Furthermore, decreasing the required window of repeated testing would enable earlier diagnostic predictions because predictions would be made using fewer (and therefore earlier) observations.

Our goal is to find a balance between accurate but costly tests and efficient but relatively inaccurate predictions. In particular, we assess and explain an individual's risk for dementia multiple years into the future using relatively easy-to-collect measures that may scale well to large aging populations. To this end, we address the following three research questions (RQs), encapsulated in Figure 1 and linked to in-text discussion. **(RQ1)** Using longitudinal clinical and cognitive data from an aging cohort study, can we effectively predict whether an individual will develop dementia? **(RQ2)** To what extent can we reduce the need for burdensome data collection while still maintaining predictive performance? We explore this question with respect to both repeated cognitive testing over multiple years and the number of required tests. **(RQ3)** Using complex models that learn interactions among features and risks, can we leverage interpretability methods to provide personalized dementia risk explanations?

Our approach makes several noteworthy contributions. First, by exploring multiple classes of ML models, we find that dementia onset (within three years) can be predicted robustly and requires cognitive measurements from *only a single ses-*

Manuscript received August 23, 2020; revised January 12, 2021; accepted February 8, 2021. Date of publication XXXXXXXXX XX, 2021; date of current version XXXXXXXXX XX, 2021. This work was supported by the National Science Foundation under Grants DBI-1759487, IIS-1901386, and IIS-1813675; The Bill & Melinda Gates Foundation under Grant INV-004841; and the National Institutes of Health under Grants R01 NIA AG 061132 and R01 LM 012810. *(* and ** indicate equal contribution.) (Corresponding author: Su-In Lee.)*

The authors are with the Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195 (e-mail: nbbwang@cs.washington.edu; amokeson@cs.washington.edu; althoff@cs.washington.edu; suinlee@cs.washington.edu)

Digital Object Identifier XXXXXXXXXXXXXX

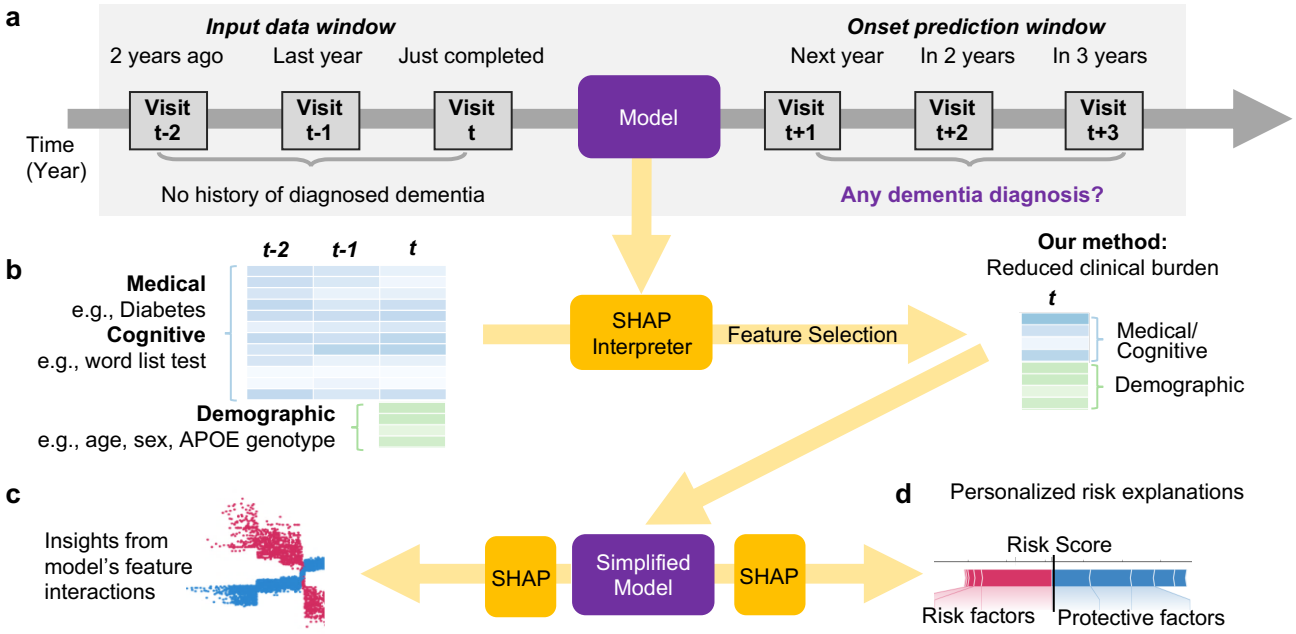

Fig. 1. Overview of our approach to producing efficient and explainable dementia onset risk predictions. We link figure components to research questions (RQs) and in-text discussion. (a) RQ1: Sections III-A and III-B. (b) RQ2: Sections III-C and III-D. (c) RQ3: Section III-E.

sion. Second, by using an interpretability method to measure sample-level feature importance, we can identify a *small subset of tests* that provide similar predictive value to a standard battery, while only taking one fifth of the time to administer. Third, each dementia risk prediction estimate is accompanied by *individual explanations of risk*, which may aid clinicians in tailoring care to their patients.

II. Related Work

State-of-the-art dementia diagnosis. Many studies have sought to predict the presence of dementia based on brain scans and other metrics. For example, deep learning has improved AD classification using both magnetic resonance imaging and positron emission tomography scans [4]–[6]. Adding lifestyle and cognitive factors has additionally improved prediction performance for AD onset [7]. Although these studies have shown success in AD diagnosis and risk prediction, neuroimaging data requires significant amounts of time and funding, making it intractable for widespread use. In contrast, we develop imminent dementia predictions based on inexpensive measures. Our approach also complements current approaches by highlighting high-risk individuals who might benefit most from more extensive testing.

Basic risk factors. Without expensive brain imaging, it is common to predict the onset of dementia from age, sex, education, and genetic factors [11]–[13]. In particular, variations in APOE, the gene encoding Apolipoprotein E protein, are thought to be the main genetic factor impacting AD risk [12], [14]. However, using only these basic risk factors produces non-robust predictions [15]. Here, we augment these primary risk predictors by adding cognitive and medical variables.

Modeling cognition trajectories. Because dementia is characterized by a rapid decline in cognitive functioning, studies have used cognitive variables to predict its onset [8]. Johnson *et al.* [9] characterized cognitive trajectories for elderly individuals with and without AD and found that precipitous drops in cognition tend to occur between one and three years prior to dementia diagnosis. Based on this result, we use up to three years of past data to predict imminent dementia onset. Unlike these longitudinal cognition studies, however, we evaluate the need for repeated testing and attempt to reduce the burden on both clinicians and participants of required study visits to achieve accurate, but efficient predictions of dementia onset.

Diagnosing cognition status. Some research has focused on assessing whether an individual already has dementia [16] or mild cognitive impairment (MCI) [17] via short questionnaires. Multiple cognitive assessments have been developed to efficiently diagnose MCI [3], [18], such as the Mini-Mental State Examination (MMSE). Further studies have used MCI diagnoses made by clinicians [19] and MMSE test scores [20] to predict future dementia onset. Building on these successes, we highlight a set of easily administered tests that significantly outperform the sole use of these clinical tests.

III. Results

We use data from the Religious Orders Study and Rush Memory and Aging Project (together known as ROSMAP) [11], [21], two longitudinal aging cohort studies, to build dementia onset risk prediction models (Section VI-A). During each yearly visit, individuals provide medical information and undergo extensive cognitive testing (Table VI). We generate samples with at least three years of consecutive visits and no dementia history and then build models to predict imminent dementia onset (i.e., a diagnosis within the next three years). Results described below are based on 9,103 samples from 1,597 individuals, split into stratified training and test sets.

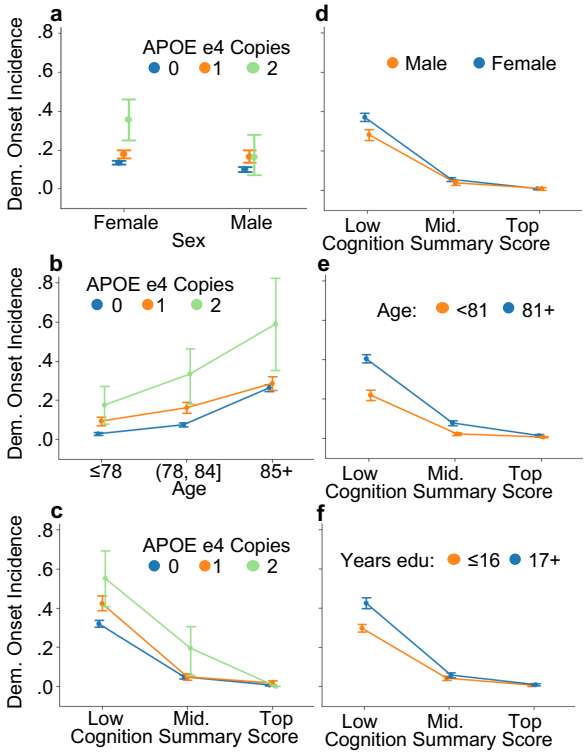

Fig. 2. Average imminent dementia onset rates (with 95% confidence intervals) by demographic and cognitive factors, highlighting non-linear and interaction effects.

A. Preliminary analyses reveal feature interactions

Preliminary data exploration comparing imminent dementia and control cases reveals many significant differences in the outcome variable among demographic and cognitive variables (Table VI). Additionally, strong correlations are seen between many features and the outcome variable, as well as among features themselves. This is expected since many of the cognitive tests assess the same cognitive domains. Together, these observations suggest that we could train an effective imminent dementia classifier from the available features. Furthermore, the high inter-relatedness of features indicates that some may provide redundant information and may therefore be reduced.

We also explore non-linear and interaction effects in our data to identify appropriate model classes. From these analyses, we observe two notable complex interactions, shown in Fig. 2. First, having a single APOE e4 allele seems to modulate dementia risk in particular groups: males (Fig. 2a), people under 85 (Fig. 2b), and relatively low cognition-scorers (Fig. 2c). We observe similar modulation among carriers of two APOE e4 alleles (e.g., females), although they represent less than 2% of our sample (Table VI). Second, we see a strong interaction between overall cognition and many demographic features. Having a high cognition score may buffer dementia risk regardless of demographic factors, while demographic features might confer more information about risk when they coincide with low cognition. For example, APOE e4-carriers (Fig. 2c), females (Fig. 2d), older individuals (Fig. 2e), and highly educated individuals (Fig. 2f) seem to exhibit especially high risk if they are also low cognition-scorers. Due to such non-linear effects among our features, a complex model may be useful for capturing interactions among features and risk.

Table I
AVERAGE CROSS-VALIDATION (CV) PERFORMANCE STATISTICS FOR EACH MODEL (± STANDARD ERROR).

Model	CV Accuracy	CV AUROC	CV AUPRC
XGB	**0.9046** ± 0.0045	0.9163 ± 0.0044	0.6763 ± 0.0132
LR	0.9045 ± 0.0048	**0.9205** ± 0.0044	**0.6893** ± 0.0110
MLP	0.9036 ± 0.0056	0.9186 ± 0.0050	0.6694 ± 0.0144
LSTM	0.9021 ± 0.0050	0.9047 ± 0.0168	0.6691 ± 0.0189

B. Multivariate models enable dementia risk prediction

To answer our first research question, we initially aim to build an ML model that can accurately predict dementia onset. To do so, we evaluate the prediction performance of multiple model classes and techniques to address *class imbalance* and *time-series data* using stratified cross validation (CV) within our training set. Due to class imbalance in our dataset (13.7% rate of dementia onset), we consider various downsampling options. Due to the data's longitudinal nature, we explore the use of time encodings to pre-process input data (e.g., moving averages; described in Section VI-D). We find that models trained without downsampling or specialized time encodings had similar or better CV accuracy, AUROC, and AUPRC scores across all model classes described below, and thus proceed with these selections for all subsequent model tuning.

For our prediction task, we compare the performance of four classes of ML models: (1) regularized logistic regression (LR), (2) XGBoost (XGB), (3) multi-layer perceptron (MLP), and (4) long short-term memory network (LSTM). For each model class, we perform extensive hyperparamter selection across five stratified cross-validation (CV) splits (within the training set). Table I shows the top-performing models in each class (Section VI-D relates tuning procedure details).

In general, we find that many of the model classes achieve similar predictive performance. MLP, LR, and XGB models perform similarly (within the standard error ranges) with respect to AUROC and AUPRC. Among complex model classes, we chose the XGB model because the neural network methods (MLP and LSTM) exhibit unstable performance, as shown by their large error bars in Fig. 3 (particularly when trained on a single year of data). We opt for an XGB final model over a linear (LR) one because: (1) Unlike linear models, XGB is able to learn non-linear and interaction effects like those found in our data. Prior meta-analyses of dementia risk prediction suggest that the linearity assumption does not hold for critical risk factors (consistent with our observations in Figure 2) [22]. Another study found that, even when producing equally accurate predictions, linear methods applied to non-linear data sets tend rely on irrelevant features [23]. Thus, XGB may learn a richer representation of the true complex relationships among features. (2) Due to the non-linear and interaction effects learned by XGB, we can obtain personalized risk explanations for each individual via interpretability methods (e.g., SHAP, Section VI-E), whereas linear models place the

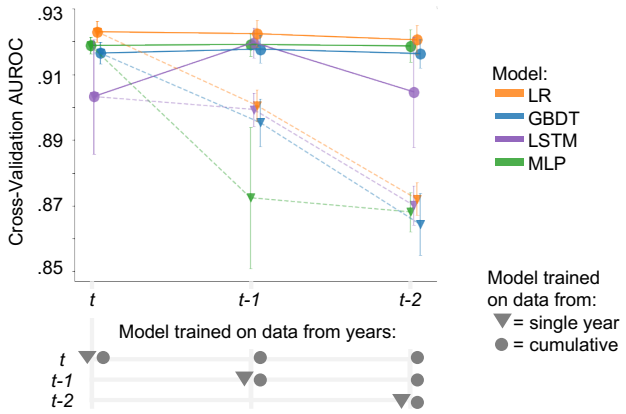

Fig. 3. Average cross-validation area under the receiver operating curve (AUROC) for our four models trained on different combinations of yearly visits. Circle marks show that cumulative data has limited value, while triangle marks highlight the importance of recent data.

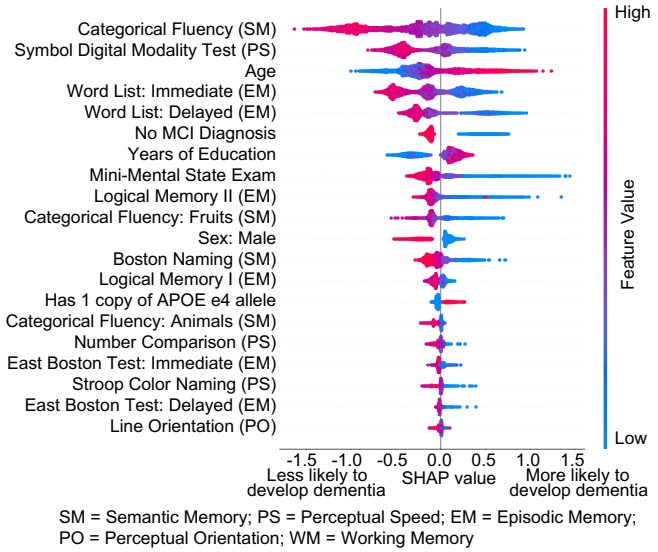

Fig. 4. SHAP summary plot: violin plot of the 20 most informative features of the XGBoost current year model, ordered by importance. Each point is a training sample colored by its feature value. The point's x-axis position is the feature's contribution to the final risk prediction.

same importance on each feature across individuals. Thus, we elect to perform final analyses with an XGB model but compare these results to a linear model for completeness.

C. Recent, not cumulative, observations are needed for effective dementia onset prediction

We answered our first research question by successfully predicting future dementia onset from three years of consecutive ROSMAP study visits. However, the use of repeated visits may be unrealistic for predicting dementia onset in the general population since only 16% of American seniors receive regular cognitive assessments [10]. Therefore, we turn to RQ2 to evaluate whether we can reduce the burden of repeated cognitive testing (i.e., do we need multiple years of cognitive measurements to make an accurate prediction?). To that end, we evaluate our model's CV AUROC when we reduce the number of consecutive visits in the inputs (Section VI-D). As we reduce the number of cumulative years the model sees during training (Fig. 3, circular markers), we find no major changes in model performance across all four model classes. This suggests that requiring multiple years of consecutive data is not necessary for accurate predictions, which may reduce the burden of regimented follow-up testing in the clinical setting.

Next, we identify the relative importance of recent data by evaluating the model trained on a single year of past data alone. As expected, we see a decline in prediction performance for models trained on older data (Fig. 3, triangular markers). Although the most effective prediction models were trained with the most recent year of observations (t), we evaluated the stability of our final conclusions by repeating analyses shown in Fig. 5 and Table II using data from $t-1$ and data from $t-2$ (the same data shown by triangle markers in Fig. 3). In both cases, models show slight drops in all performance metrics, but our models outperform baselines by similar margins on time $t-1$ and $t-2$ data compared to t data. Together, these results imply that recent cognitive measurement are vital for predicting imminent dementia status, but that repeated testing is not needed for accurate dementia onset prediction since recent data may supersede outdated cognitive information obtained in past years.

Finally, we apply SHAP [24], a local feature attribution method, to our XGB model trained on all three years of consecutive data to ascertain whether the model relies on previously collected data (see Sections III-D and VI-E). We find that the model's top ten features consist of demographic data or tests from the most recent year: even when provided access to measurements from prior years, our model still tends to focus on more recent data. Based on these CV results, we decided to train the final models using only the current year (t) of data, a decision that enables earlier, more efficient predictions that need not wait for additional years of cognitive tests before generating a dementia prediction.

D. Efficient and effective dementia onset predictions can be made with a small subset of features

After extensive cross-validation experiments, we settle on a final XGB model using the hyperparameters selected based on CV performance. This final "All Features" XGB model is trained on all training data using all available features from year t only. Table II shows held-out test set performance metrics for this model. To drive further insights, we use SHAP local feature explanations [24] to interpret the final model. To see which features our XGB model relies on, we aggregate the local explanations of our training samples to obtain global insights (Section VI-E). Fig. 4 shows the top 20 most important features (ranked by their average SHAP importance magnitude across all samples).

First, we note that the feature attributions are consistent with findings in the literature, validating our modeling approach and SHAP interpretations. For example, nearly all previous work [11] has found females, older individuals, and carriers of an APOE e4 allele to be at higher risk of dementia, consistent

Table II
TEST PERFORMANCE OF FINAL MODELS (± STANDARD ERROR FROM BOOTSTRAP RE-SAMPLING).

	Test Accuracy	**Test AUROC**	**Test AUPRC**	**Relative IDI** (Simplified with APOE vs. row)
Final models				
All Features (XGB)	**0.8975** $\pm$ 0.0002	**0.8977** $\pm$ 0.0003	**0.6387** $\pm$ 0.0010	-0.0571
Simplified (with APOE) (XGB)	0.8947 $\pm$ 0.0002	0.8903 $\pm$ 0.0003	0.6236 $\pm$ 0.0010	–
Simplified (no APOE) (XGB)	0.8964 $\pm$ 0.0002	0.8896 $\pm$ 0.0003	0.6184 $\pm$ 0.0010	0.0084
Baseline models in our study				
Linear Selected Features (LR)	0.8825 $\pm$ 0.0002	0.8224 $\pm$ 0.0004	0.4907 $\pm$ 0.0011	0.6012
Linear Selected Features (XGB)	0.8781 $\pm$ 0.0002	0.8050 $\pm$ 0.0005	0.4771 $\pm$ 0.0011	0.7432
Baseline feature sets in the literature				
Demographics + MCI (XGB) [19]	0.8770 $\pm$ 0.0002	0.8203 $\pm$ 0.0005	0.4449 $\pm$ 0.0011	0.5058
Normalized Cognitive Features Sum (LR)	0.8737 $\pm$ 0.0002	0.8128 $\pm$ 0.0005	0.4473 $\pm$ 0.0011	0.9804
Demographics + MMSE30 (XGB) [20]	0.8748 $\pm$ 0.0002	0.8124 $\pm$ 0.0005	0.4273 $\pm$ 0.0011	0.6707
Demographics (XGB) [11]	0.8593 $\pm$ 0.0003	0.7215 $\pm$ 0.0005	0.2660 $\pm$ 0.0008	2.9291

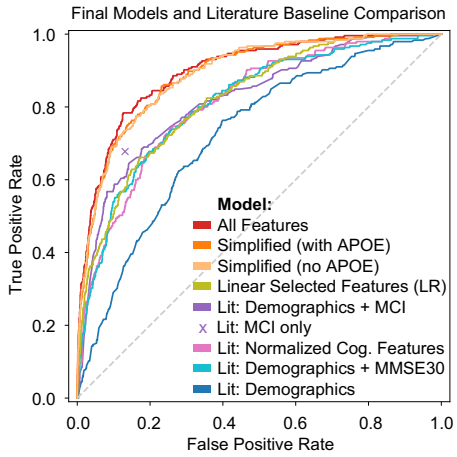

Fig. 5. Receiver operating curves: final models and baselines from the literature (Lit). Area statistics in Table II.

with Fig. 4 SHAP explanations. Similarly, as expected, low performance on all cognitive tests contributes to a higher risk score. In contrast, our years of education feature attributions are not consistent with the literature (which find negative associations between high education and dementia incidence). We discuss this result further in Section IV (Limitations).

As we move down the list of top-ranked features, we see a dramatic drop in the magnitude of SHAP values (i.e., relative influence of a given feature on the final prediction). We therefore hypothesized that future dementia onset can be predicted using only the most informative features. For evaluation, we choose the top four demographic features and top four cognitive tests and use them to train a simplified prediction model. The top demographic features (age, sex, education, and APOE genotype) are widely cited as being important [11] and are simple to measure. The four top cognitive tests chosen are: categorical fluency (Cat Flu; 2 minutes; semantic memory); symbol digit modality test (SDMT, $\leq$ 5 minutes, perceptual speed); word list test (WL, 3 minutes, episodic memory), and mini-mental state exam (MMSE30, 5-10 minutes, general cognition); Table III describes these tests. Interestingly, each test lies in a different cognitive domain in Table VI, indicating that the model relies on diverse and non-redundant cognitive attributes. From our simplified feature set, we train two "simplified" final models on our full training set: one including and one excluding the APOE genotype (which, though commonly used in prior studies, is not always available in clinical settings). Although cognitive diagnostic status (MCI diagnosis) was ranked among the top influential features, we excluded it from our simplified models: it is very time consuming to obtain in the ROSMAP study (since it is based on all cognitive tests and a clinician examination), and it may be difficult to obtain in the general aging population.

Unlike the above use of SHAP-based feature selection from our XGB model, feature selection for linear models involves choosing those with the highest magnitude regression coefficients. For comparison with our SHAP selection method above, we use standard feature selection based on the final LR model's coefficient magnitudes. For consistency, we use the same four demographic features as above and then select the cognitive features with the highest-magnitude regression coefficients: digits forward, digits backward, digits ordering (all working memory), and the East Boston Test (episodic memory) (21 minutes total; See Table III).

We compare our final simplified XGB models to the XGB model trained on the full feature set, an LR and XGB model trained using the features from linear feature selection, and a baseline XGB trained on multiple commonly used clinical baseline feature sets (Section VI-F). Fig. 5 shows the held-out test set's receiver operating curves for all models, highlighting the sensitivity and specificity based on all decision boundaries on the test set. Using a decision cut-off of 0.5, we report true negatives, false positives, false negatives, and true positives for our top models and the top performing baseline model in Table IV. For each model, Table II lists the area under the receiver operating curves (AUROC), precision recall curves (AUPRC), and the accuracy at a 0.5 decision cut-off point. Additionally, we calculate the relative integrated discrimination improvement (IDI), comparing the "Simplified (with APOE)" model's discrimination ability to every other model [25].

Fig. 5 and Table II show that our methods significantly outperform XGB models trained with more restricted feature sets documented in the literature [19], [20]. Notably, our simplified dementia onset predictor is only slightly less accurate than the model trained using all features (with the non-APOE model showing only a slight performance drop compared to the APOE model). Additionally, our SHAP approach selects

Table III

SELECTED COGNITIVE TESTS FROM XGBOOST (XGB) AND LINEAR REGRESSION (LR) MODELS (COGNITIVE DOMAINS SHOWN IN TABLE VI). FULL COGNITIVE BATTERY: 98 MINUTES.

Test (Domain)	Time (min)	Description
Selected cognitive tests from XGB model:		
Categorical fluency (SM)	2	Subject names as many items in a category as they can in a minute (Rounds: animals, fruits).
Symbol digit modality (PS)	$\leq$5	Subject learns a symbol-to-digit mapping, then must substitute digits when symbols are shown.
Word list (EM)	3	Subject hears a list of 10 words, then is tested on immediate recall, delayed recall, and recognition (selecting correct words from distractors).
Mini-mental state exam	$\leq$10	Short diagnostic general cognition test for dementia.
Selected cognitive tests from LR model:		
Digits Forward (WM)	5	Given a list of numbers, subject repeats them in the same order as given.
Digit Ordering (WM)	5	Given a list of numbers, subject repeats them in numerical order.
East Boston Test (EM)	6	After hearing a short story, subject recalls story units immediately and after distractor-filled delay.
Digits backward (WM)	5	Given a list of numbers, subject repeats them in the reverse order as given.

Table IV

USING A 0.5 DECISION CUT-OFF, WE REPORT THE NUMBER OF TRUE NEGATIVES (TN), FALSE POSITIVES (FP), FALSE NEGATIVES (FN) AND TRUE POSITIVES (TP) IN THE TEST SET.

	TN	FP	FN	TP
All Features (XGB)	1510	50	135	110
Simplified (with APOE) (XGB)	1513	47	143	102
Demographics + MCI (XGB) [19]	1479	81	141	104

Table V

CROSS-STUDY TEST SET PERFORMANCE FOR ROS VS. MAP MODELS.

		AUROC for Test Set Samples	
		ROS (N=1156)	MAP (N=649)
Training Samples	ROS (N=4506)	0.8848	0.8948
	MAP (N=2792)	0.8792	0.8851

a more effective set of features than the classic linear feature selection approach, further supporting our choice of using a non-linear model (i.e., XGBoost).

Together, these results show that computing SHAP feature importances for our XGB model allows us to identify measures that are particularly useful in our model and thus dramatically improve prediction performance over more basic clinical baselines by including a few short cognitive tests. These tests are standardized and simple to administer; any primary care physician or assistant could conduct them during a patient's annual physical exam (taking a total of 15-20 minutes to administer compared to the 98 minutes required for all tests in the ROSMAP neuropsychological battery in Table III).

Cross-cohort generalizability. Due to differences in study design and measured features, it is uncommon for dementia prediction studies to validate findings with external datasets [15], [22]. While our data is comprised of pooled ROS and MAP samples, the studies recruit participants from different groups (clergy from Catholic religious organizations across the US and individuals in retirement facilities throughout northern Illinois, respectively) [11], [21], and these studies differ in demographic and lifestyle factors and outcomes (see Section VI-C). Thus, we seek to evaluate the cross-study generalizability of our final models. Using our previously defined training and test splits, we retrain our Simplified (with APOE) model separately for ROS and MAP training samples and evaluate each model's performance separately for ROS and MAP held-out test samples. In both cases, the "external" and "internal" test set AUROCs are within 0.01 of each other (Table V). Furthermore, similar tests would be selected if we were to perform feature selection based on models trained separately from each cohort (the same top four tests for ROS, and three of four top tests–with the number comparison test replacing MMSE–for MAP). Together, these findings indicate that the model generalizes stably and effectively across cohorts, both in terms of predictive performance and selected features.

Missing data experiments. As shown in Table VI, many features have missing values and are imputed for all analyses (Section VI-C). In particular, some features are missing at significantly different rates for control versus dementia onset cases. To ensure that our promising results were not driven by a confounding effect of imputing features at different rates between case and control groups, we experiment with removing potentially confounded samples and features as follows. We first exclude features with one-fifth of samples missing (Stroop color naming and Stroop word reading tests). We next exclude all samples with a missing observation for any of the remaining 10 features with significantly different rates of missingness between control and dementia onset cases, resulting in a new dataset with 8,392 samples (92% of the original dataset). First, we note that final model performance on this filtered dataset (test AUROC=0.8952) is very similar to performance from the full dataset (test AUROC=0.8977). Importantly, our SHAP feature rankings (generated via average SHAP importance magnitude) result in the same top four selected cognitive tests as the original dataset. Further, we observe similar performance for the final simplified model (test AUROCs of 0.8865 and 0.8903, respectively, for filtered and original datasets). Together, these experiments indicate that imputing missing features had little effect on our final models.

E. SHAP provides personalized risk explanations

We turn now to our third research question, which addresses personalized dementia risk explanations. Because XGB learns complex relationships among features (unlike linear models), we examine SHAP interaction values among pairs of features (Section VI-G). For example, according to XGB interactions, having one copy of the APOE e4 allele impacts an individual's XGB risk prediction, particularly if he or she has a low cognition score (Fig. 6a, consistent with Fig. 2c) or is younger (Fig. 6b, consistent with Fig. 2b). Finally, males, especially

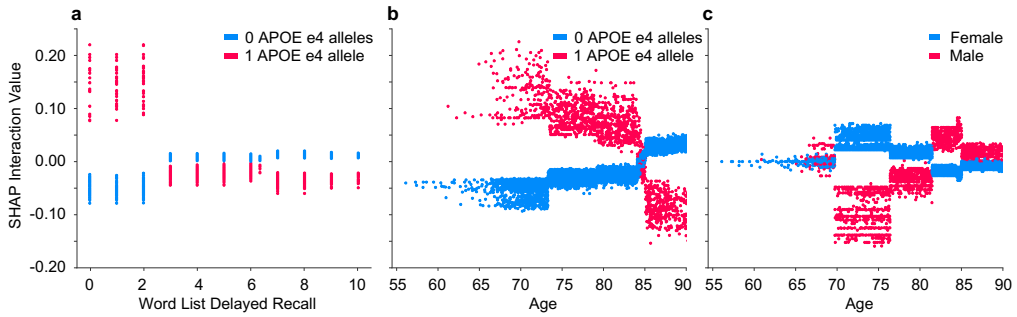

Fig. 6. SHAP interaction values for selected pairs of features in our final Simplified (with APOE) XGBoost model.

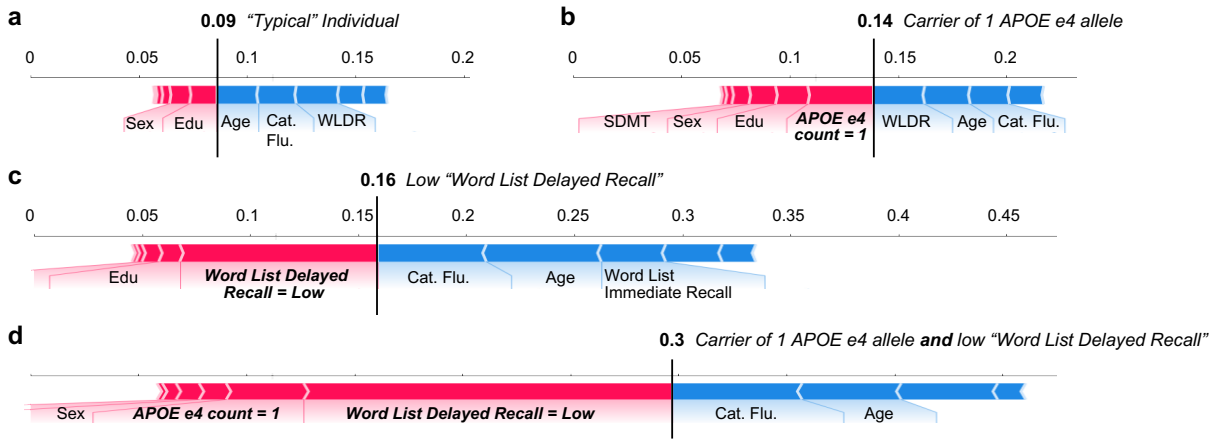

Fig. 7. Feature explanations for synthetic samples: (a) risk and explanations for a "typical individual" in the ROSMAP data, (b, c) perturbations to single features (bolded), (d) the combined effects of both risk factors.

those younger than 80, are at particularly low risk for developing dementia (Fig. 6c, consistent with earlier findings [11]). Thus, by using SHAP to interpret our simplified XGB model, we find that aggregating non-linear feature effects across samples reveals relevant interactions learned by the model, and that these interactions are consistent with our data's structure (Fig. 2) and prior literature.

Beyond receiving a risk score, using XGB and SHAP feature attributions gives patients and their medical practitioners a personalized explanation of risk (i.e., how particular features drive the XGB's prediction). To illustrate how this benefit works, we generate a synthetic sample that represents the "typical sample" in our dataset (with mode- or average-valued features) and display the risk score and explanation in Fig. 7a. We show perturbations to single features of APOE (where we change the APOE e4 allele count from zero to one) in Fig. 7b and the word list delayed recall (WLDR) score from the average value to two standard deviations below the average in Fig. 7c. In both examples, we see that the perturbed variable becomes the primary risk factor driving up the risk score compared to the "typical individual."

Finally, in Fig. 7d, the effect of both risk factors from parts b and c shows that the combined risk of having one APOE e4 allele and a low WLDR score substantially increases risk. In particular, the jump in risk from both risk factors (a 0.21 increase over the "typical individual") far exceeds the additive effects of each single risk score alone (0.05 and 0.07 increase, respectively). A linear model, in contrast, would have produced additive predictive importance values and therefore would have failed to identify a compounding effect of these features. This example highlights the ability of our XGB model with SHAP interpretations to provide presonalized risk explanations based on a combination of feature values. This ability may prove to be powerful in clinical settings because it would help clinicians discuss the unique configuration of risk factors relevant to individual patients.

IV. Discussion

Comparison with previous findings. Reviews of dementia prediction studies have found that combinations of cognitive tests have aided in the prediction of dementia onset [15], [26]. In particular, for predicting conversion from MCI to dementia, combining episodic memory tests with executive functioning or language tests tended to produce high predictive accuracy [26]. A review of community-based aging cohort studies (consistent with our approach) also found that using three or four tests spanning multiple cognitive domains led to improved predictions of dementia onset for 2.5 to 5 year follow-ups [15]. Compared to our results, these studies reported similar or lower AUROCs (ranging from 0.83 to 0.88); however, each study was based on samples from different cohorts (ranging from 478 to 551 total participants) and with different follow-up periods, so direct comparison may not be appropriate. Importantly, despite being performed on a

larger cohort (1,597 individuals) and using a non-linear XGB model (unlike the previous studies, which all relied on linear analyses), our approach identified a small number of tests spanning multiple cognitive domains (Table III) as predictors of dementia, consistent with these prior studies.

Longitudinal input data. Curiously, our analyses show the modest value of longitudinal measurements. Because dementia is an acquired condition marked by cognitive decline, one might expect to see gradual changes in cognition prior to dementia onset. In fact, our choice of a three-year input data window was based on observed cognitive changes preceding dementia in prodromal cases [9]. However, because our goal is to predict a *future* dementia diagnosis (not a current one), changes in cognition scores may be less useful than expected. Our results seem consistent with other studies, which reported limited value for cognitive changes in predicting future dementia onset. In particular, one study found that reliable change indices (RCIs) for MMSE had low predictive accuracy for dementia onset [27]. Furthermore, because longitudinal input data inherently requires rarer datasets with multiple cognitive assessments, RCI-based studies have often failed to achieve the same predictive accuracy as single-observation studies [15].

Limitations. Our final dataset contained 9,103 samples from 1,597 individuals, of which, 521 developed dementia. Although our study is based on a larger dataset than prior studies mentioned above [15], future studies should replicate our findings in other populations. Because we rely on samples from the ROS and MAP cohorts, our findings are subject to potential bias introduced by each cohort's procedures. In particular, for our ROSMAP samples, approximately three quarters come from females and two thirds from participants with 16 or more years of education. The unusually high education levels in our data may explain why some feature explanations for education level are inconsistent with findings in the literature. Future studies should especially explore sex- and education-based dementia risk in a more balanced dataset.

It is uncommon for dementia prediction studies to validate their findings externally due to prohibitive differences in study design, populations, and measured features [15]. According to a recent review [22], less than a quarter of examined ML studies externally validated their findings (the majority of which were imaging studies with harmonized measurements). As with many prior studies, we could not directly assess our findings on an external dataset. Nevertheless, despite differences between the ROS and MAP cohorts, our models generalized well between them when trained separately.

Additionally, the Stroop color naming and word reading tests had high levels of missingness in our dataset (Table VI), so it is possible that those tests may have been more highly ranked if they were observed in more samples. However, most features had relatively low rates of missingness (Table VI), and we found that there was not a significant relationship between feature missing rates and their SHAP importance for our final XGB model (Pearson correlation $r = -0.11, p = 0.46$). Furthermore, analyses described in Section III-D showed that our imputation methods did not significantly affect our findings.

Finally, our initial choice of time window (three input years and three years of onset monitoring) limited the samples that were included from the ROSMAP dataset, biasing our sample against individuals with fewer than six yearly visits. We made this decision based on prior work, which suggested at least one-to-three years of cognitive data are useful for modeling cognitive decline in prodromal dementia patients [9]. We also viewed this as a necessary drawback in order to evaluate whether longitudinal data is needed for accurate prediction.

V. Conclusion

We conducted an in-depth analysis of many ML models, sampling techniques, and usages of time-series data to obtain models that predict imminent dementia onset more accurately than basic demographics-based or single-test approaches and more efficiently than prediction from a full neuropscyhological battery. Importantly, we can accurately predict imminent dementia diagnoses using data from just one clinical visit consisting of only demographic information and four easily measured cognitive tests that can be conducted in less than 20 minutes (five times shorter than the standard cognitive battery in the ROSMAP study). By using complex non-linear models and leveraging ML interpretability methods, we also generate personalized explanations of risk predictions that account for non-linear and interaction effects. These findings may provide substantial clinical value given the growing aging population and low rates of routine medical assessments. Our method could be scaled to explain and highlight at-risk individuals for additional dementia screenings, preventative treatments (when they become available), and enable planning for a potential imminent diagnosis. Our study takes important steps toward using complex models to generate explainable dementia risk predictions from relatively cheap metrics. While our findings highlight the effectiveness of our approach, more studies are needed to to provide further validation for use in clinical practice. Nevertheless, we provide a framework with which others may replicate our experiments and construct models tailored to other cohorts and their measured cognitive tests.

VI. Methods

We now describe in detail how we produced the results described in this paper. Additionally, our code for reproducing these results is available at: github.com/suinleelab/EEDRP.

A. Dataset

The Religious Orders Study (ROS) [11] and Memory Aging Project (MAP) [21] are complementary epidemiological studies that each enroll persons without dementia who agree to annual evaluations and eventual organ donation. ROS enrolls clergy living communally from 40 Catholic groups across the US (primarily employed or retired nuns, priests, and brothers). This study group was selected because communal living provided both high follow-up rates and relative consistency in life experiences and socioeconomic factors. However, as a volunteer cohort of Catholic clergy, the samples are not representative of a wider population of eldery individuals [11]. As a complementary study, MAP recruited participants from a wider range of life experiences throughout northeastern Illinois. Participants are primarily enrolled from continuous

care retirement communities (ranging in care levels from independent living to nursing on campus). To reduce participant burden and facilitate high follow-up rates, data was collected via home visits. Clinical data collection procedures were consistent between both studies to allow the data to be merged for analyses [21]. Due to their recruitment strategies, follow-up rates of survivors reached around 95% for both studies. Compared to ROS samples, MAP samples were obtained from relatively fewer males (23.6% vs 31.9%) and from individuals who were older (83 vs 80 years on average) and less educated (15 vs 18 years on average). MAP samples also had higher rates of MCI (21.2% vs 19%) and a higher incidence of dementia onset within three years (15.3% vs. 12.7%).

Upon entering the study, participants share demographic information (e.g., sex, age) and blood samples for genotyping. At each yearly visit, they provide updated medical information and undergo a battery of cognitive tests, resulting in repeatedly measured variables. We predict dementia onset from 41 separate variables (per time point; note that categorical variables were one-hot encoded, leading to 49 total features), which we list in Table VI. In total, the data contains 3,194 individuals with one to 23 annual visits. Of all participants with at least two years of visit data and no original dementia diagnosis, 619 (23.7%) were eventually diagnosed with dementia.

B. Data processing: generating samples

Our prediction task (Fig. 1) is: Using data from his/her three most recent practitioner visits, does an individual with no history of dementia experience dementia onset within the next three years? In particular, our selected time-frame was based on prior findings that a precipitous drop in cognitive abilities is usually observed one-to-three years prior to a dementia diagnosis [9]. To construct the appropriate dataset, we narrow our analyses from the 3,194 existing participants to 1,597 individuals for whom we have enough observations.

Many participants had more than six consecutive yearly visits, so we applied a sliding window of six years over their available consecutive visits, thereby generating at least one sample, but often more, per participant. Each sample is split into an input window (consisting of the first three consecutive visits $t-2, t-1$, and t) and onset prediction window (consisting of the next three consecutive visits: $t+1, t+2$, and $t+3$), as illustrated by positive (dementia onset) and negative (no dementia onset) examples in Fig. 8. Because the goal is to predict future dementia onset in individuals who do not yet have dementia, we exclude all samples in which dementia is already present during visits $t-2, t-1$, and t (e.g., Fig. 8, Example Participant A, samples 2 and 3). Finally, we applied sliding windows of four and five years to identify any additional positive onset cases (e.g., Fig. 8, Participant B, Samples 4 and 5), which helped to mitigate our class imbalance. This procedure could not be used to find negative dementia onset samples because all three future years must be known to definitively rule out a dementia diagnosis.

C. Data processing: pre-processing for all models

After combining all valid six-year windows (and four- and five- year windows where appropriate), we have a sample size

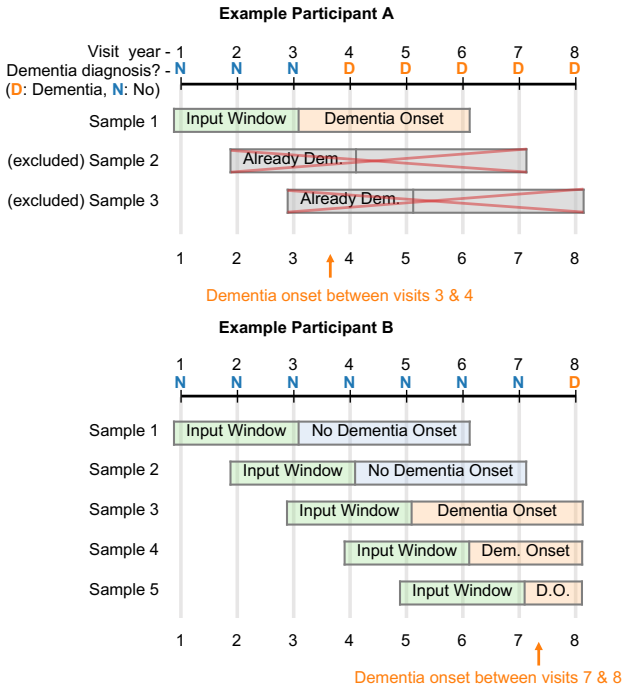

Fig. 8. Examples of samples from sliding windows. Our samples have no history of dementia during the first 3 years, and either no onset for *all* of the next 3 years (negative case) or a dementia diagnosis in *any* of the next 3 years (positive case).

of 9,103 samples, of which 13.7% were labeled as positive dementia onset cases (derived from 1,597 individuals, of which 521 developed dementia). For each model next described, our model inputs consist of variables obtained during the first three visits ($t-2, t-1$, and/or t), called the input data window, and the outputs are a prediction of whether the individual was diagnosed with dementia at any of visits $t+1, t+2$, or $t+3$. Table VI shows all demographic, cognitive, and medical features from our 9,103 samples (at time t), split by dementia onset label, and associated between-group differences.

Since some downstream analyses require variables to be on the same scale, we standardize all continuous variables for our input data and use z-scores as features for all models. To maintain consistent scores across time points, z-scores are calculated based on time t observations (and the same re-scaling procedure based on time t is applied to observations at $t-1$ and $t-2$). For categorical variables, we apply one-hot encoding. We note in Table VI that most variables have some missing observations across our samples. We impute all missing samples using the mean for continuous variables and the mode for categorical variables (across all samples). Using chi-square tests of independence, we find that some cognitive tests have significantly different missingness rates between dementia onset and control groups, although the rates tend to be low (between 0.2% and 6%, except for Stroop test variables). Additional analyses described in Section III-D confirm that the effects of imputing values did not impact our final results (compared with filtering out the affected cases).

We next describe model selection with cross-validation and then evaluation on a test set. For each analysis, we use the

Table VI

BETWEEN-GROUP BASELINE (TIME t) STATISTICS. WE PROVIDE SUMMARY STATISTICS FOR EACH GROUP (INCLUDING MISSINGNESS RATES AND INDICATORS OF SIGNIFICANTLY HIGHER RATES OF MISSINGNESS FOR ONE GROUP). (*<.05, $^{**}p < .01$, $^{***}p < .001$ FOR STATISTICAL TESTS.)

	All samples: Between-group test statistic	**Summary Statistics** (% samples missing and associated significance)	
		Controls: No impending dementia ($N = 7866$)	**Dementia onset within 3 years** ($N = 1244$)
Demographics			
Age	$t = -29.60^{***}$	80.12 ± 6.77 (0%)	86.19 ± 6.37 (0%)
Sex: % male	$\chi^2 = 15.43^{***}$	29.5% (0%)	24.0% (0%)
Years of education	$t = 1.45$	16.91 ± 3.61 (0.1%)	16.75 ± 3.56 (0.2%)
Race (White/Black/Native American/Asian)	$\chi^2 = 12.72^{**}$	94.0%/5.5%/0.3%/0.2% (0%)	94.2%/4.9%/0.2%/0.6% (0%)
Ethnicity: % Hispanic	$\chi^2 = 0.10$	3.1% (0%)	3.3% (0%)
# APOE e4 copies (0/1/2)	$\chi^2 = 54.59^{***}$	78.6%/20.3%/1.1% (1.5%)	70.3%/27.0%/2.8% (1.4%)
Episodic Memory (EM)			
Word list: immediate (1min)	$t = 40.10^{***}$	20.57 ± 4.55 (2.4%)	15.01 ± 4.08 (2.3%)
Word list: delayed (1min)	$t = 45.43^{***}$	6.76 ± 2.16 (2.4%)	3.69 ± 2.34 (2.4%)
Word list: recognition (1min)	$t = 32.82^{***}$	9.85 ± 0.56 (2.3%)	9.02 ± 1.74 (2.7%)
East Boston test: delayed (3min)	$t = 26.25^{***}$	9.89 ± 1.73 (0.3%)	8.46 ± 2.03 (1.0%**)
East Boston test: immediate (3min)	$t = 34.59^{***}$	9.64 ± 1.90 (0.5%)	7.41 ± 3.07 (1.2%**)
Logical memory I (3min)	$t = 37.73^{***}$	14.34 ± 4.10 (2.3%)	9.51 ± 4.44 (2.1%)
Logical memory II (3min)	$t = 40.47^{***}$	13.23 ± 4.45 (2.4%)	7.60 ± 4.81 (2.4%)
Perceptual Orientation (PO)			
Line orientation (15min)	$t = 13.54^{***}$	10.59 ± 2.97 (3.8%)	9.32 ± 3.02 (6.1%***)
Progressive matrices (20min)	$t = 22.82^{***}$	11.65 ± 2.82 (5.1%)	9.61 ± 2.79 (8.2%***)
Perceptual Speed (PS)			
Symbol digits modality test (5min)	$t = 37.91^{***}$	41.77 ± 10.09 (3.9%)	29.72 ± 9.87 (7.2%***)
Number comparison (3min)	$t = 26.12^{***}$	26.22 ± 7.23 (3.7%)	20.29 ± 7.12 (6.2%***)
Stroop color naming (3min)	$t = 22.30^{***}$	20.19 ± 7.34 (65.2%***)	12.34 ± 6.55 (60.0%)
Stroop word reading (3min)	$t = 13.66^{***}$	48.87 ± 13.53 (65.3%***)	39.74 ± 14.55 (60.1%)
Semantic Memory (SM)			
Boston naming (5min)	$t = 29.15^{***}$	14.19 ± 0.98 (3.0%)	13.22 ± 1.51 (4.0%)
Categorical fluency: animals (1min)	$t = 33.21^{***}$	18.25 ± 5.45 (0.1%)	12.90 ± 3.96 (0.4%)
Categorical fluency: fruits (1min)	$t = 37.98^{***}$	18.26 ± 5.13 (0.2%)	12.44 ± 4.15 (0.6%*)
Categorical fluency (combined)	$t = 40.00^{***}$	36.51 ± 9.42 (0.1%)	25.33 ± 7.08 (0.4%)
National adult reading test (2min)	$t = 5.15^{***}$	8.49 ± 1.94 (3.6%)	8.17 ± 2.14 (6.7%***)
Working Memory (WM)			
Digits backward (5min)	$t = 16.94^{***}$	6.61 ± 2.05 (0.4%)	5.56 ± 1.82 (0.9%*)
Digits forward (5min)	$t = 11.92^{***}$	8.43 ± 1.98 (0.2%)	7.70 ± 1.99 (0.6%)
Digit ordering (5min)	$t = 21.30^{***}$	7.60 ± 1.56 (1.0%)	6.57 ± 1.67 (2.4%***)
Global Cognition			
Mini-mental state exam (5-10min)	$t = 45.16^{***}$	28.59 ± 1.51 (2.2%)	26.20 ± 2.71 (1.4%)
Medical history/lifestyle factors			
MCI (No/Yes/Yes-other)	$\chi^2 = 1685.26^{***}$	86.9%/12.8%/0.3% (0%)	37.1%/60.7%/2.3% (0%)
Medical conditions sum	$t = -3.77^{***}$	1.68 ± 1.16 (2.0%)	1.82 ± 1.21 (2.0%)
Vascular disease burden	$t = -7.02^{***}$	0.45 ± 0.66 (2.0%)	0.59 ± 0.75 (2.0%)
Vascular disease risk	$t = -1.31$	0.87 ± 0.81 (1.2%)	0.90 ± 0.77 (1.0%)
Any history of:			
cancer	$\chi^2 = 2.48$	40.2% (2.0%)	37.8% (2.0%)
claudication	$\chi^2 = 19.62^{***}$	22.4% (2.0%)	28.2% (2.0%)
diabetes	$\chi^2 = 0.44$	11.6% (2.0%)	12.3% (2.1%)
diabetes medication	$\chi^2 = 1.97$	15.6% (1.2%)	17.2% (1.0%)
head injury with loss of consc.	$\chi^2 = 0.03$	9.7% (2.0%)	9.8% (2.0%)
heart disease	$\chi^2 = 10.04^{**}$	12.7% (2.0%)	16.0% (2.0%)
hypertension	$\chi^2 = 7.24^{**}$	56.9% (2.0%)	61.0% (2.0%)
stroke	$\chi^2 = 36.05^{***}$	9.7% (0.9%)	15.3% (0.5%)
thyroid disease	$\chi^2 = 1.51$	24.1% (2.0%)	25.8% (2.0%)

same stratified training and test sets. To avoid contaminating our test set with training examples, we split our data by participants so that all samples from a single individual fall into the training set or test set, but not both. Of our 1,597 participants, we assigned one fifth of them to the test set (1,805 associated samples) and the remaining individuals (7,298 associated samples) to our training set. Next, we randomly divide our training set participants into five stratified cross-validation splits. All splits were performed in a stratified manner to maintain consistent ratios of AD to control cases.

D. Building and evaluating prediction models

We evaluated modeling options under several domains: sampling techniques to address class imbalance, time encoding techniques, and model class. Our modeling choices were based on average accuracy, areas under the receiver operating curve (AUROC), and areas under the precision recall curve (AUPRC) across five cross-validation (CV) folds.

Downsampling. The dataset has a class imbalance of 13.7% positive labels since few individuals experience dementia onset in any given 3-year window. Therefore, we experimented with four different downsampling techniques: (1) no downsampling, (2) class re-weighting (incorporated into loss functions during

Table VII
ENCODING METHODS USED FOR TIME-SERIES FEATURES.

Name	Description
All data	Unaltered data from t, t-1, t-2
Moving averages	(1) Unaltered features from t, (2) One simple moving average feature derived from t, t-1, and t-2 features, and (3) Three exponential moving average features with half-life values of 1, 2, and 3 years derived from t-2 features
Slopes	(1) Unaltered features from t, (2) the change in features from each year to the next (i.e., $v_t - v_{t-1}$ and $v_{t-1} - v_{t-2}$ for variable v), and (3) the overall change in scores from the earliest year to the current year (i.e., $v_t - v_{t-2}$)

training), (3) random downsampling (randomly selecting as many negative as positive samples), and (4) matched pairs downsampling. In (4), for each positive sample, we select the closest negative sample based on sex, age, and education (greedily, without replacement). Due to equal or better prediction performance across five-fold CV, all final models are trained with no downsampling (see Section III-B).

Time-series encoding. Because of the longitudinal nature of many features, we evaluated methods for incorporating repeatedly observed variables: (1) all data (no special encoding), (2) moving averages, and (3) slopes (see Table VII). Per Section III-B, training with all data yielded similar or better CV performance, and thus was used for all subsequent models.

Model Type. We compared the performance of four classes of ML models: (1) logistic regression (LR; implemented with Scikit-Learn [28]), (2) gradient-boosted decision trees via the XGBoost algorithm (XGB; known for handling mixed feature types and medical data well [29]), (3) multi-layer perceptrons (MLP; deep learning approach), and (4) long short-term memory networks (LSTM; time series aware deep learning approach). Both deep learning approaches were implemented in Keras [30] and tensorflow [31]. For each model class, we evaluated several hyperparameter settings and selected the setting with the highest average CV AUROC (reported in Table I). We share our final hyperparameters, along with average CV performance across modeling choices described in this section, in our code repository: github.com/suinleelab/EEDRP.

Training with fewer input years. We next evaluate whether we can reduce the burden of repeated cognitive testing (i.e., do we need multiple years of data to accurately predict dementia?). We compare performance of models trained on the last 3 year's visits with models trained on fewer time points: the last 2 years' visits (t and $t-1$) and the most recent visit (t) (circular markers in Fig. 3). We also evaluate the importance of recent data for impending dementia predictions: in addition to evaluating the model trained on the most recent visit alone (t), we also train models on data from single visits one and two years earlier ($t-1$ alone and $t-2$ alone) (triangular markers in Fig. 3). Results (Section III-C) indicate that recent, but not repeated, measurements are needed for accurate prediction.

To further explore whether the model relies on past data, we perform feature importance analysis using SHAP (Section VI-E) on our XGB model trained on the last 3 years of data. The model's top ten features are from time t (including demographic features), which provides further evidence that relying on past measurements is not necessary.

E. Model interpretation with SHAP explanations

To explore what the model is learning and drive further insights, we use SHAP local feature explanations applied to our XGB model (trained on the full feature set with current year, t, data). To obtain global feature importances, we aggregate local feature attributions across training samples. Features with higher global importances have more impact on model predictions across samples (Fig. 4). Next, we select a subset of available features based on their global SHAP ranking: the top 4 demographic features (age, sex, education, APOE genotype) and the top 4 cognitive tests (with their subtests; Table III). Our final feature set excludes the variable "No cognitive impairment diagnosis" because it is a cognitive diagnosis that is inefficient to obtain (based on both the full 98-minute neuropsychological battery and a medical review from a physician). Finally, to compare feature selection using SHAP global importances to the more typical global feature selection method in linear models, we use the same demographic features and select the four cognitive tests with the highest-magnitude coefficients from the linear model (Table III).

F. Measuring final model performance

First, we compare final test performance of XGB trained on the full feature set compared with 2 simplified feature sets: (1) the top four demographic features and the top 4 cognitive tests, and (2) the same set of features but excluding APOE genotype, which may be expensive to obtain for those without existing genotype data. To compare selected features from SHAP to those from a simple linear method, we also report performance for XGB and LR models trained on the features selected via LR coefficients, described above (Table III).

Finally, for comparison with our methods, we also generate multiple baseline XGB models trained on features commonly used as risk indicators in the literature (Fig. 5 and Table II): (1) demographic features (above) [11]; (2) MCI diagnosis and demographic variables [19]; (3) the MMSE30 and demographic variables [20]; and (4) the sum over all normalized cognitive test scores controlled for age, sex, and education.

Fig. 5 displays ROC curves, showing the performance of models at all possible decision cut-off points. We also show confusion matrices for the top-performing baseline and final models using an example cut-off of 0.5 (Table IV). Table II summarizes all performance metrics, including confidence intervals from bootstrap resampling of the test set (repeated 1,000 times). Per Fig. 5 and Table II, the features selected from the XGB model result in similar AUROCs compared with the full feature set (and outperform the linearly selected features). While the full cognitive battery requires 98 minutes of cognitive testing, we achieve similar predictive value using only four tests that take under 20 minutes.

G. Examining SHAP explanations in the final model

Feature interactions. To explore the complex interactions learned by the XGB model, we examine SHAP interaction values among pairs of features in our final simplified model. For each sample in our training set, the SHAP interaction value

for two features represents the remaining combined feature effect after removing individual main effects of both features.

Fig. 6 shows feature interactions in the XGB model: each point is a training sample colored by one feature and placed on the x-axis according to its value for the second feature. The y-axis indicates the sample's SHAP interaction value (refer for more detail to [23]). In parts b and c, samples with ages over 90 were censored due to privacy requirements. Higher absolute value y-axis values in these plots indicate that the XGB model makes risk predictions based on feature combinations rather than independently based on single features.

Personalized explanations. For any sample, we can generate a SHAP force plot to explore personalized risk explanations provided by SHAP applied to our final XGB model [32] (e.g., Fig. 7). These plots indicate both the model's dementia onset risk prediction and the SHAP values for the highest-contributing features impacting the prediction (pink arrows for risk factors, and blue arrows for protective ones).

To clarify the variations in explanations in a controlled setting, we generate four synthetic examples. First, we show a SHAP force plot for a "typical individual" in our dataset (i.e., a sample with mean or mode values for all features; Fig. 7a). A "typical individual" has a low risk of developing dementia in the next three years since the diagnostic rate for dementia is low in any 3-year period. Next, we show perturbations to single feature values for APOE (where we change the APOE e4 allele count from zero to one; Fig. 7b) and word list delayed recall (WLDR) score (from the mean value to two standard deviations below the mean, i.e., from six words remembered to just one; Fig. 7c). Finally, we simultaneously perturb both risk factors above and show that the combined risk of having both one APOE e4 allele and a low WLDR score leads to a large, non-linear jump in risk that exceeds the combined single effects of each feature alone (Fig. 7d).

REFERENCES

[1] "2020 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 16, no. 3, pp. 391–460, 2020.

[2] M. Prince, R. Bryce, and C. Ferri, "World alzheimer report 2011: The benefits of early diagnosis and intervention." Alzheimer's Disease International, 2011.

[3] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-mental state: A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189 – 198, 1975.

[4] H.-I. Suk and D. Shen, "Deep learning-based feature representation for ad/mci classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Nagoya, Japan, 2013, pp. 583–590.

[5] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[6] R. Cui and M. Liu, "Rnn-based longitudinal analysis for diagnosis of alzheimer's disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1 – 10, 2019.

[7] D. E. Barnes, K. E. Covinsky, R. A. Whitmer, L. H. Kuller, O. L. Lopez, and K. Yaffe, "Predicting risk of dementia in older adults: The late-life dementia risk index," *Neurology*, vol. 73, no. 3, pp. 173–179, 2009.

[8] S. Lee *et al.*, "Episodic memory performance in a multi-ethnic longitudinal study of 13,037 elderly," *PloS one*, vol. 13, no. 11, p. e0206803, 2018.

[9] D. K. Johnson, M. Storandt, J. C. Morris, and J. E. Galvin, "Longitudinal study of the transition from healthy aging to alzheimer disease," *Archives of neurology*, vol. 66, no. 10, pp. 1254–1259, 2009.

[10] J. Gaugler, B. James, T. Johnson, A. Marin, and J. Weuve, "2019 alzheimer's disease facts and figures," *Alzheimers & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[11] D. Bennett, J. A Schneider, Z. Arvanitakis, and R. S Wilson, "Overview and findings from the religious orders study," *Current Alzheimer Research*, vol. 9, no. 6, pp. 628–645, 2012.

[12] V. Chouraki *et al.*, "Evaluation of a genetic risk score to improve risk prediction for alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 53, no. 3, pp. 921–932, 2016.

[13] A. C. Naj *et al.*, "Age-at-onset in late onset alzheimer disease is modified by multiple genetic loci," *JAMA neurology*, vol. 71, no. 11, p. 1394, 2014.

[14] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy," *Nature Reviews Neurology*, vol. 9, no. 2, p. 106, 2013.

[15] B. C. Stephan, T. Kurth, F. E. Matthews, C. Brayne, and C. Dufouil, "Dementia risk prediction in the population: are screening models accurate?" *Nature Reviews Neurology*, vol. 6, no. 6, pp. 318–326, 2010.

[16] Z. Arabi, S. A. S. A. Rahman, H. Hazmi, and N. Hamdin, "Reliability and construct validity of the early dementia questionnaire (edq)," *BMC geriatrics*, vol. 16, no. 1, p. 202, 2016.

[17] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[18] Z. S. Nasreddine *et al.*, "The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, pp. 695–699, 2005.

[19] A. Bozoki, B. Giordani, J. L. Heidebrink, S. Berent, and N. L. Foster, "Mild Cognitive Impairments Predict Dementia in Nondemented Elderly Patients With Memory Loss," *Archives of Neurology*, vol. 58, no. 3, pp. 411–416, 03 2001.

[20] D. B. Hogan and E. M. Ebly, "Predicting who will develop dementia in a cohort of canadian seniors," *Canadian Journal of Neurological Sciences*, vol. 27, no. 1, p. 18–24, 2000.

[21] D. Bennett, J. A Schneider, A. S Buchman, L. L Barnes, P. A Boyle, and R. S Wilson, "Overview and findings from the rush memory and aging project," *Current Alzheimer Research*, vol. 9, no. 6, pp. 646–663, 2012.

[22] J. Goerdten, I. Cukic, S. O. Danso, I. Carriere, and G. Muniz-Terrera, "Statistical methods for dementia risk prediction and recommendations for future work: A systematic review," *Alzheimer's Dementia: Translational Research Clinical Interventions*, vol. 5, pp. 563–569, 2019.

[23] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable ai for trees." *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 2017, pp. 4765–4774.

[25] E. W. Steyerberg *et al.*, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology*, vol. 21, no. 1, p. 128, 2010.

[26] S. Belleville *et al.*, "Neuropsychological measures that predict progression from mild cognitive impairment to alzheimer's type dementia in older adults: a systematic review and meta-analysis." *Neuropsychology Review*, vol. 27, pp. 328–353, 2017.

[27] A. Hensel, T. Luck, M. Luppa, H. Glaesmer, M. C. Angermeyer, and S. G. Riedel-Heller, "Does a reliable decline in mini mental state examination total score predict dementia?" *Dementia and geriatric cognitive disorders*, vol. 27, no. 1, pp. 50–58, 2009.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[30] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[31] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

[32] S. M. Lundberg, *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature biomedical engineering*, vol. 2, no. 10, pp. 749–760, 2018.