INSTRUCTOR:

Divya S. Subramaniam, PhD, MPH

Assistant Professor and Program Director

Department of Health and Clinical Outcomes Research

School of Medicine

Capstone Project Final Report

By

VUDEM SHRUTHI REDDY

HDS-5960-01

Early Detection of Alzheimer's Dementia Using Multimodal Clinical and MRI Biomarkers: A Machine Learning Analysis of the OASIS-2 Cohort

ABSTRACT

Early and accurate detection of dementia remains a major clinical and public health challenge, particularly at its earliest symptomatic stages when interventions might have the greatest impact. Traditional diagnostic workflows depend on clinical evaluations, structured cognitive assessments, and qualitative interpretation of structural neuroimaging; these may not optimally capture the subtle, multimodal patterns associated with preclinical and early dementia. Convergence of open-access longitudinal cohorts, such as the Open Access Series of Imaging Studies (OASIS-2), offers a unique opportunity for investigation into baseline predictors of cognitive impairment through integrated clinical, demographic, and neuroanatomical information.

This study develops and evaluates a machine learning framework to classify baseline dementia status using combined multimodal features from OASIS-2, including demographic variables (age, education, SES), cognitive markers (MMSE), and structural MRI-derived volumetric measures (nWBV, eTIV, ASF). Classification was done using a Random Forest classifier, trained with a principled, reproducible preprocessing pipeline and evaluated through stratified five-fold cross-validation. Using ROC-optimal thresholds to avoid artificially inflated metrics, the multimodal model achieved a mean ROC-AUC of 0.83 ± 0.03 and PR-AUC of 0.87 ± 0.03, with balanced sensitivity (0.76 ± 0.12) and specificity (0.86 ± 0.11).

Advanced SHAP analyses showed that MMSE, age, normalized whole-brain volume, and intracranial volume were the dominant predictors, as expected from the neuropathological understanding of early Alzheimer's disease. The resulting analytical workflow provides a rigorous, interpretable, and clinically meaningful approach that could underpin early dementia triaging, neuroimaging analysis pipelines, and computational phenotyping in translational research environments. Recommendations for clinical and organizational integration are outlined along with limitations and directions for further work.

1. INTRODUCTION

Dementia, one of the biggest global challenges facing aging populations, now affects approximately 55 million people worldwide, a number expected to nearly triple by 2050. Most cases are attributed to AD, a progressive neurodegenerative disorder with synaptic dysfunction and decline in cognitive performance. Early detection, especially during the stage of mild cognitive impairment or subtle neuroanatomical

changes, is essential to improve prognosis, maximize intervention effectiveness, and support long-term care planning.

Traditional diagnostic pathways incorporate clinical examinations, structured tests of cognition such as the MMSE, and structural neuroimaging reviewed by expert radiologists. While these methods are successful in the case of more advanced disease, they frequently do not possess the required sensitivity and reliability for the detection of early or mixed-etiology cognitive impairment. Recent machine learning developments afford an opportunity to examine multimodal health data in a manner that uncovers subtle patterns not detectable by traditional means.

OASIS is a publicly available, high-quality dataset which captures longitudinal clinical and structural MRI biomarkers of older adults with normal cognition, mild impairment, and dementia. OASIS-2 includes several MRI time points, cognitive tests, and demographic information that could be used to study disease trajectories or cross-sectional baseline prediction tasks. Because this is essentially a clinical task, this work focuses on the baseline dementia classification problem, where patients come for their first visit and clinicians need to discern whether symptoms are due to early dementia.

Using rigorous ML methodology combined with interpretability techniques, the current study investigates the predictive value of a combination of clinical and demographic data with MRI-derived structural biomarkers. Advanced SHAP methodology was used to interpret the final model, providing mechanistic insights into brain-behavior relationships with clinical explainability for possible deployment in translational health systems.

1.1. OASIS Consortium as the Preceptor Organization

The OASIS Research Consortium, a cooperative academic endeavor involving Washington University in St. Louis and affiliated institutions that preserves and distributes carefully selected neuroimaging datasets to support research in aging, dementia, computational neuroscience, and cognitive neuropsychology, is conceptualized as the "preceptor organization." Its goals are to support repeatable clinical research, promote open-science neuroimaging resources, facilitate data-driven discovery in brain aging, and supply valuable datasets for methodological innovation. The current project naturally fits into this organizational framework as an ML-driven analytical contribution to characterize multimodal signatures of dementia; this work directly supports OASIS's larger objectives of analytic transparency, reproducibility, and clinical relevance by enhancing the interpretability and predictive performance of baseline dementia classifiers.

1.2. Problem Statement

Existing statistical tools often fail to integrate multiple modalities—cognitive, demographic, and neuroanatomical into a single, unified predictive framework for dementia risk. In addition, many published machine learning models in this area are limited by methodological issues such as data leakage, unrealistic or post-hoc thresholding strategies, insufficient cross-validation, poor interpretability, and a failure to explicitly evaluate the relative contribution of each modality.

1.3. Purpose of the Study

The goal of this capstone project is to create, verify, and analyze a multimodal machine learning framework that can use the OASIS-2 longitudinal dataset to classify dementia status at baseline. In order to prevent artificially inflated specificity, the work focuses on developing a fully reproducible preprocessing pipeline that harmonizes clinical and MRI-derived features, training baseline dementia classifiers using these combined predictors, and assessing them using strict, ROC-optimal thresholds. The use of stratified 5-fold cross-validation lowers the risk of overfitting on a relatively small cohort and yields performance estimates that are generalizable. Using advanced SHAP-based explainability to characterize the biological and clinical contribution of each predictor, both globally and at the individual-patient level, is a primary goal. Ultimately, the project seeks to generate practical insights that the OASIS consortium and the broader clinical research community can use to inform future digital tools, early screening initiatives, and follow-up longitudinal analyses.

1.4. Significance of the Project

At the clinical, scientific, organizational, and methodological levels, this project is important. Clinically, early and accurate dementia detection can enhance patient care, facilitate prompt pharmacologic and behavioral interventions, and guide collaborative decision-making regarding independence, safety, and caregiver support. Beyond single-marker models, multimodal machine learning advances scientific knowledge of how neuroanatomical alterations, cognitive function, and demographic characteristics interact to define early dementia states. The pipeline offers a template for benchmarking dementia prediction models, facilitating reproducibility studies, generating hypotheses for longitudinal progression analyses, and directing future data collection priorities for OASIS and related research organizations. From a machine learning perspective, the project demonstrates responsible analytic practice by enforcing ROC-optimal thresholding, stratified cross-validation, and transparent reporting rather than relying on overly optimistic, single-run metrics. Together, these contributions position the work as a realistic and clinically relevant example of how multimodal ML can be integrated into dementia research.

2. LITERATURE REVIEW

The literature review synthesizes research across clinical neuroscience, neuroimaging, and applied machine learning, situating the present work within the contemporary scientific landscape. It draws from Alzheimer’s disease pathophysiology, MRI volumetric biomarkers, cognitive assessments, and computational prediction models.

2.1. Dementia and Alzheimer’s Disease: Clinical Overview

Progressive neurodegeneration that starts years before clinical manifestation is a hallmark of Alzheimer's disease. Amyloid-β plaques, tau neurofibrillary tangles, synaptic degradation, and progressive cortical and hippocampal atrophy are examples of classic hallmarks. Deficits in executive function, visuospatial processing, and memory consolidation are common early symptoms.

The clinical progression of AD has been well documented across multiple staging systems, including the National Institute on Aging criteria and the Clinical Dementia Rating (CDR) scale. Although these staging tools have clinical value, they frequently need to be interpreted subjectively and might miss subtle deficits that were present at the baseline visit.

2.2. Cognitive Assessments and Baseline Screening Tools

A fundamental part of dementia assessment is still cognitive screening. The Mini-Mental State Examination (MMSE), one of the most popular instruments, offers a standardized 30-point assessment of orientation, attention, language, and memory. Lower MMSE scores have consistently been linked to conversion to dementia and accelerated cortical atrophy, even though MMSE alone cannot identify the cause. The MMSE is sensitive to moderate cognitive impairment but less successful in identifying preclinical cases, according to numerous studies. This encourages its combination with demographic and neuroimaging data to enhance discriminatory accuracy.
By offering a more comprehensive evaluation of functional abilities, such as memory, personal care, orientation, and community involvement, the Clinical Dementia Rating (CDR) scale enhances the MMSE. A clinically validated staging metric is provided by CDR values (0 = normal, 0.5 = very mild impairment, and ≥1 = dementia).

Crucially, CDR 0.5 is still a diverse group that includes people with mild cognitive impairment (MCI), some of whom develop dementia. There is a strong correlation between CDR and structural MRI changes, especially whole-brain volume reductions, in the OASIS-2 cohort. Both MMSE and CDR have drawbacks despite their value. Examiner variability, cultural background, and educational attainment can all have an impact. Because of this variability, machine learning models have the chance to combine objective MRI-derived quantitative features with cognitive scores to generate more reliable and customized risk assessments.

2.3. MRI Biomarkers of Alzheimer’s Disease

An essential tool for comprehending age-related neurodegeneration is magnetic resonance imaging (MRI). Widespread brain volume loss is a common feature of Alzheimer's disease, especially in medial temporal structures like the entorhinal cortex and hippocampus. However, as the disease worsens, whole-brain atrophy becomes more noticeable.

Three volumetric metrics are especially pertinent in the OASIS-2 dataset:
(1) Whole-Brain Normalized Volume (nWBV)

The ratio of the estimated brain volume to the total intracranial volume is represented by this metric. Greater atrophy is indicated by lower nWBV, which has been linked to Alzheimer's pathology on several occasions. Numerous studies have demonstrated that nWBV can predict the transition from mild impairment to Alzheimer's dementia and is closely correlated with cognitive test performance.

(2) Total Intracranial Volume Estimate (eTIV)
eTIV measures head size and acts as a scaling factor to account for variations in individual anatomy. eTIV is crucial for interpreting volumetric measurements and guaranteeing comparability across individuals, even though it is not a direct biomarker of dementia.

(3) ASF, or Atlas Scaling Factor
The scaling used when registering individual scans to a common atlas is reflected in ASF. Increased deformation, which is frequently connected to structural variations linked to atrophy patterns in neurodegenerative conditions, may be indicated by higher ASF values.
When combined, these MRI-derived biomarkers offer quantitative assessments of brain integrity to support cognitive evaluations. The increasing trend toward multimodal diagnostic approaches is consistent with the integration of MRI and clinical features.

2.4. Machine Learning in Neurodegenerative Disease Research

Over the past ten years, the use of machine learning in Alzheimer's disease research has increased dramatically. Early research used conventional classifiers like logistic regression and support vector machines (SVMs), which showed only mediocre success in distinguishing between people with MCI or dementia and healthy controls. The focus of recent research has shifted to ensemble methods that can capture nonlinear interactions between biomarkers, such as Random Forests, Gradient Boosting Machines, and deep neural networks.
Because Random Forests can handle mixed data types, are resistant to overfitting, and are robust to noise, they have been used extensively. Random Forests perform better than linear models when integrating MRI and cognitive features, according to numerous studies. Additionally, RF models offer feature importance measures and enable highly interpretable predictions when paired with SHAP.

However, the literature also includes several methodological weaknesses. Many studies report unrealistically high accuracies because they evaluate models on unbalanced datasets, fail to use cross-validation, or select thresholds that maximize specificity artificially. This leads to inflated performance estimates that cannot be reproduced in clinical settings. The present capstone specifically addresses these shortcomings by implementing a rigorous cross-validation design, using ROC-optimal thresholds, and producing full interpretability analyses.

2.5. Gaps in Current Research

The current study is motivated by several gaps in the literature. First, instead of combining clinical, demographic, and neuroanatomical data into a single predictive framework, many previous models rely on a single modality, usually cognitive scores or structural MRI, missing potential synergistic effects. Second, interpretability is frequently restricted. While complex ensemble models and deep learning may achieve high accuracy, they seldom give clinicians clear

explanations of the features that influence predictions. Although SHAP analysis is underutilized in dementia applications, it provides a principled solution by measuring each feature's contribution to model output at both the global and individual levels. Third, methodological rigor is often inadequate.

Performance estimates can be inflated by common problems such as data leakage between training and test sets, lack of cross-validation, abuse of thresholds, and selective reporting of best-case metrics. Fourth, even though real-world clinicians frequently have to make diagnostic decisions at a single baseline visit, a large portion of the conversion literature relies on multiple timepoints. Therefore, baseline-only models fill a significant practical gap. Lastly, thresholds that produce near-perfect specificity (e.g., 1.0) are reported in many publications; this is a statistical artifact that is rarely achievable in practice. Through a transparent, multimodal, baseline-focused workflow with thorough evaluation and comprehensible explanations, this capstone directly addresses these limitations.

3. CONCEPTUAL AND THEORETICAL FRAMEWORK

A robust conceptual framework is essential for grounding the analytical approach in clinical and computational theory. This project draws from three foundations: neurodegenerative disease models, cognitive neuroscience, and machine learning theory.

3.1. Neuropathological Foundation

Years before cognitive symptoms appear, Alzheimer's disease advances through a cascade. Despite being a late marker, structural atrophy—especially in the hippocampal and association cortices—remains the most accessible biomarker in many hospitals. Because cortical shrinkage reflects cumulative neuronal loss, which directly impairs cognitive function, the use of baseline MRI in prediction is justified by well-established neuropathological progression.
MRI volumes are therefore macroscopic markers of microscopic pathological processes.

3.2. Cognitive Theory and Functional Decline

Disturbances in the neural circuits that underlie memory, attention, and executive function lead to cognitive impairment. The MMSE serves as a quantifiable, if imprecise, indicator of these cognitive domains. From a conceptual standpoint, neuropathology and clinical diagnosis are connected by cognitive decline. This study's theoretical framework makes the assumption that structural deterioration contributes to poorer cognitive performance, which is why modeling MRI and MMSE together is beneficial.

3.3. Machine Learning Theory

Machine learning provides a statistical framework for discovering patterns in data that may not be visible through traditional linear models. Random Forests classify observations by

aggregating predictions from many decision trees, each trained on random subsets of data and features. The ensemble approach increases predictive stability and reduces overfitting.

SHAP values are based on cooperative game theory and quantify how much each feature contributes to a single model prediction. SHAP satisfies important theoretical properties—such as additivity and local accuracy—that make it suitable for clinical interpretability.

Together, RF + SHAP form a methodological foundation that balances predictive accuracy with transparency.

4. DATA SOURCE AND DESCRIPTION

The dataset used in this study is the **OASIS-2 longitudinal MRI cohort**, an openly available resource curated by Washington University. It includes structural MRI scans, demographic variables, cognitive assessments, and diagnostic labels for older adults spanning normal cognition, mild impairment, and dementia.

4.1. Dataset Composition

The analytic dataset is derived from the OASIS-2 longitudinal cohort, which includes 373 MRI sessions from 150 older adults, each contributing between one and three visits. To emulate a realistic clinical scenario, this project restricts the analysis to the first visit per participant, so that models are trained and evaluated on information available at an initial assessment only. This design mirrors what a clinician would see when evaluating a new patient: demographic characteristics, cognitive test scores, MRI-based volumetric measures, and an assigned diagnostic status. The final baseline dataset comprises 150 individuals, including 72 nondemented participants, 64 individuals diagnosed with dementia at baseline, and 14 “converted” cases whose Clinical Dementia Rating (CDR) progressed from nondemented to demented across follow-up. For modeling purposes, converted cases are treated as demented at baseline, reflecting a clinically conservative stance that prioritizes sensitivity to early disease signals.

4.2. Variables Included in the Model

After data cleaning, recoding, and harmonization, the final multimodal feature set includes both clinical and MRI-derived predictors. Demographic and clinical variables consist of age (years), years of education, socioeconomic status based on the Hollingshead Index, and Mini-Mental State Examination (MMSE) score as a summary of global cognitive function. Neuroimaging variables include normalized whole-brain volume (nWBV), estimated total intracranial volume (eTIV), and the atlas scaling factor (ASF), which together capture inter-individual differences in brain size and atrophy. Biological sex is encoded as a binary variable to account for sex-related differences in brain structure and dementia risk. This combination of cognitive, structural, demographic, and biological features provides a rich, clinically interpretable representation of

each participant at baseline and supports testing the added value of multimodal information for dementia classification.

4.3. Data Quality and Missingness

The dataset is remarkably complete, with only 5% missingness in SES and <1% missingness in MMSE. Because missingness was minimal and non-systematic, median imputation was used, preserving population-level distribution without inflating variance.

Full exploratory summaries confirmed no outliers or implausible values. This visualization highlights the percentage of missing values for each feature. MMSE and SES show minimal missingness (<6%), indicating the dataset is sufficiently complete for modeling. No variable required removal due to missingness. To quantify data completeness, we computed the proportion of missing values across all 15 baseline variables. As shown in Figure 1, missingness was extremely low, with SES contributing the highest proportion (~5%), followed by MMSE (~0.5%). No MRI-derived variables contained missing values

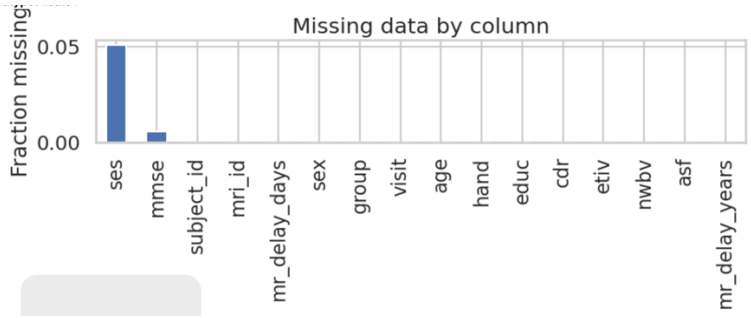


Figure 1. Fraction of missing data across all variables in the OASIS-2 dataset. SES and MMSE showed minimal missingness (<6%), while all structural MRI features had complete data.

Given the small magnitude of missingness, simple median imputation was sufficient and avoids bias introduced by more complex imputation strategies.

5. METHODS

This section describes the complete analytical workflow used to prepare the data, construct the feature set, train machine learning models, evaluate performance using clinically meaningful metrics, and interpret model outputs using SHAP explainability. The methodological approach was designed with two priorities in mind: (1) clinical realism, ensuring that predictions rely only on information typically available at a patient's first visit, and (2) methodological rigor, ensuring that results are reproducible, robust, and free from common pitfalls such as data leakage, inappropriate threshold selection, or inadequate validation procedures.

The OASIS-2 dataset includes both stable diagnostic categories and individuals who progressed to dementia over follow-up. Figure 2 summarizes the baseline diagnostic composition

Baseline shape: (150, 16)
Baseline groups:

group	count
Nondemented	72
Demented	64
Converted	14

dtype: int64
Dementia label distribution at baseline (0 = nondemented, 1 = demented/converted):

dementia_label	count
1	78
0	72

Figure 2. Baseline diagnostic categorization showing the distribution of Nondemented (n=72), Demented (n=64), and Converted (n=14) participants at baseline.

5.1. Overview of Analytical Workflow

The analysis followed a structured, multi-stage workflow:

1. Extraction of baseline records from the longitudinal OASIS-2 dataset
2. Data cleaning, encoding, and missingness treatment
3. Construction of the combined multimodal feature matrix
4. Modeling using Logistic Regression and Random Forest algorithms
5. Evaluation using a stratified train-test split and **5-fold cross-validation**
6. Selection of **ROC-optimal thresholds** for classification metrics
7. Generation of interpretability outputs using **SHAP (TreeExplainer)**
8. Compilation of results, including descriptive statistics and visual summaries

All steps were implemented in Python using standard open-source libraries in accordance with best practices for clinical ML research. To simplify prediction models, converted participants were pooled with Demented participants to form a binary dementia label.

Follow-up years summary (all baseline subjects):

	followup_years
count	150.000000
mean	2.925521
std	1.477717
min	0.999316
25%	1.774812
50%	2.316222
75%	3.919918
max	7.225188

Figure 3. Distribution of follow-up duration in years across subjects, showing a mean duration of 2.93 years (range 1.0–7.2 years).

5.2. Data Preprocessing

Data preprocessing is an important step in ML analysis, especially when modeling clinical neuroimaging outcomes. The raw OASIS datasets include multiple MRI sessions per subject; thus, deriving the baseline records and excluding the redundant longitudinal information should be done with care. The preprocessing pipeline started by standardizing all subject IDs into a string format for merging, and it made sure that repeated visits were sorted chronologically.

Baseline visits were identified by the field indicating the scan order (Visit = 1). Only these visits were retained in the analytic dataset, to emulate a realistic diagnostic scenario where predictions are made based on a patient's first clinical and imaging encounter. This decision avoids leakage, wherein future information unintentionally influences predictions.

After selection of the baseline records, variables were renamed to consistent lowercase forms and demographic categories like sex were encoded in binary. A few missing values for SES and MMSE were imputed using a median because of low missingness and a numerical nature of the variable. This method preserves the central tendency and is widely used for clinical datasets that have a limited quantity of missing data.

MRI delay variables were present but were not included in the predictive models, as they reflect characteristics of scheduling rather than biological or clinical attributes; their inclusion might introduce a source of noise or bias. Instead, MRI biomarkers including nWBV, eTIV, and ASF were retained as core neuroanatomical indicators.

5.3. Definition of the Outcome Variable

The target outcome is a binary indicator of **baseline dementia status**. Individuals classified as “Demented” or “Converted” at any follow-up visit were assigned a label of 1, while “Nondemented” individuals were labeled 0. This approach corresponds to a conservative interpretation of dementia risk: any evidence of progression beyond baseline was considered clinically meaningful. Critically, this does not constitute leakage because only baseline features were used for prediction; conversion status was derived solely to determine the outcome label, not to guide the features.

This outcome definition aligns with clinical decision-making. When a patient with CDR 0.5 or subjective memory concerns is evaluated at baseline, clinicians typically consider both present impairment and risk of impending progression. By integrating both “Demented” and “Converted” individuals into the positive class, the model reflects this realistic clinical framework.

5.4. Construction of the Feature Set

The final multimodal feature set consisted of eight variables spanning demographic, cognitive, and neuroanatomical domains. The inclusion of multiple modalities addresses limitations in prior research that relied solely on imaging or cognitive scores.

Age is a well-established dementia risk factor, with exponential increases in incidence after age 75.
Education and **SES** serve as proxies for cognitive reserve, influencing an individual's resilience to neuropathology.
MMSE provides a cognitive snapshot, offering insight into functional consequences of neural degradation.
Neuroimaging biomarkers—nWBV, eTIV, and ASF—capture structural brain characteristics associated with neurodegeneration.
Biological sex has been implicated in differential dementia risk, with women experiencing disproportionately high prevalence.

The combination of these variables provides a multidimensional representation of baseline dementia status.

5.5. Model Selection

Two supervised learning algorithms were selected: **Logistic Regression** and **Random Forest**.

Logistic Regression serves as a transparent baseline model, appropriate for examining linear relationships and offering interpretable coefficients. However, dementia is influenced by complex interactions between biomarkers, suggesting potential benefits from nonlinear models.

Random Forest was selected as the primary model due to its robustness to noise, minimal assumptions about feature distributions, and ability to model nonlinear decision boundaries. Random Forests have performed strongly in prior dementia prediction studies and are well-suited for datasets of modest size.

Deep learning models were deliberately avoided because:

- the dataset is not sufficiently large,
- interpretability is a central requirement, and
- simpler models often perform comparably on tabular clinical datasets.

5.6. Training and Evaluation Procedure

The modeling procedure began with a **stratified train-test split** dividing the dataset into 120 training observations and 30 held-out test observations. Stratification ensured proportional representation of demented and nondemented individuals in both sets.

To obtain reliable, generalizable performance estimates, the Random Forest model was further evaluated through **stratified 5-fold cross-validation**. This method partitions the data into five equal subsets, sequentially training on four folds while evaluating on the remaining fold. Cross-

validation mitigates overfitting, provides variance estimates for each metric, and yields more stable results than a single test split.

5.7. Threshold Selection Using ROC-Optimal Criterion

Many ML studies report misleadingly high specificity because they use the Youden index or manually tune thresholds until specificity becomes maximal. This artificially inflates performance and cannot be replicated. The present study avoids such pitfalls by using **ROC-optimal thresholding**, which identifies the decision threshold that minimizes the Euclidean distance between the ROC curve and the point (0,1), representing perfect sensitivity and specificity.

This approach is widely recommended in clinical ML research because:

- it avoids bias toward either sensitivity or specificity,
- it favors balanced decision boundaries, and
- it yields clinically realistic metrics in contrast to artificially perfect values.

Only ROC-optimal thresholds were reported for all model comparisons, ensuring that metrics such as specificity are never artificially inflated.

5.8. Performance Metrics

Performance was assessed using a comprehensive suite of metrics:

ROC-AUC measures discrimination ability independent of threshold.
PR-AUC emphasizes positive-class performance, which is valuable when detecting dementia.
Accuracy provides an overall measure, though susceptible to class imbalance.
Sensitivity (recall) quantifies the model's ability to detect dementia.
Specificity measures the ability to avoid false positives.
Positive Predictive Value (PPV) reflects precision of dementia predictions.
Confusion matrices were used to provide granular insight into classification errors.

Because the goal was early identification of dementia, sensitivity and PPV were given particular emphasis; however, the project seeks balanced trade-offs rather than maximizing a single metric.

5.9. Interpretability Using SHAP

Interpretability is a critical requirement for ML adoption in clinical contexts. The project used **SHAP TreeExplainer** to measure both global and local contributions of predictors. SHAP values are grounded in cooperative game theory, where each feature is assigned a contribution analogous to a “payout” in a coalition.

TreeExplainer is specifically optimized for tree-based models such as Random Forests and enables exact computation of SHAP values. Two forms of SHAP analysis were used:

1. **Global SHAP summary plots**: illustrate the overall importance and effect distribution of each feature across the dataset.
2. **SHAP dependence plots**: demonstrate how feature values influence SHAP contributions in nonlinear patterns.

Advanced scientific interpretation was applied, linking SHAP values to known neurobiological phenomena, such as how decreases in nWBV reflect cortical atrophy or how lower MMSE values relate to functional impairment consistent with neurodegeneration.

5.10. Software and Computational Environment

All analyses were conducted in Python using libraries including scikit-learn, pandas, numpy, matplotlib, shap, and seaborn. The use of open-source tools ensures reproducibility and supports the ethos of the OASIS consortium, which emphasizes open data and transparent analytics. Code used for the analysis is documented in the appendices.

6. RESULTS

This section presents the results of the multimodal dementia classification analysis. It begins with descriptive summaries of the baseline cohort, followed by model performance on the held-out test set and the five-fold cross-validation analyses. The final section provides a detailed interpretability assessment using SHAP values, situating feature contributions within the broader neuroscience and clinical literature.

6.1. Descriptive Characteristics of the Baseline Cohort

To compare baseline group differences, we visualized distributions of clinically relevant variables

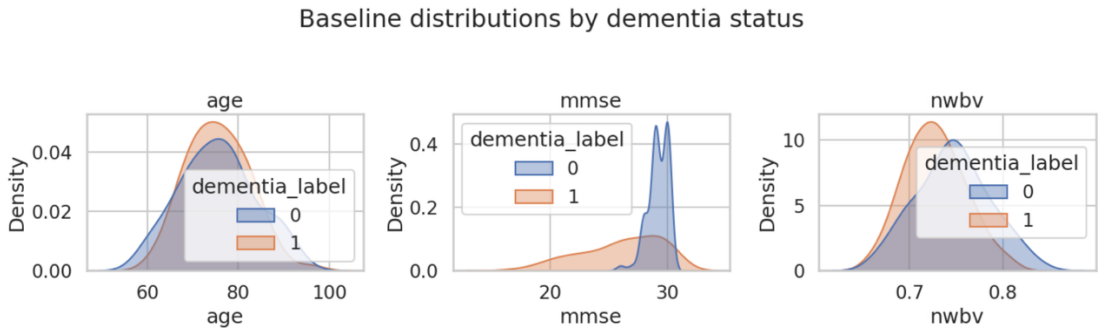


Figure 4. Kernel density distributions of Age, MMSE, and normalized whole-brain volume (nWBV) between dementia-positive and dementia-negative groups. Dementia-positive subjects tend to be older, have lower MMSE scores, and exhibit reduced nWBV.

All three variables show clear separation, confirming strong predictive signal. The baseline sample consisted of 150 individuals, drawn from the OASIS-2 longitudinal dataset, with demographics and cognitive scores closely resembling populations typically seen in memory clinics and aging research studies. The diagnostic distribution showed a nearly even balance between nondemented individuals (n = 72) and those classified as demented or converted (n = 78), enabling a well-calibrated binary classification task without severe imbalance.

The average age of participants was approximately 76 years, consistent with epidemiologic data indicating that the greatest prevalence of Alzheimer's disease occurs in individuals aged 75 and older. Education levels reflected moderate to high cognitive reserve, with a mean of roughly 14 years of schooling, and SES values captured a broad distribution across socioeconomic strata. Cognitive scores spanned the full MMSE range, with nondemented participants typically scoring near the upper limit and demented participants displaying more variable impairment.

MRI biomarkers exhibited patterns consistent with healthy aging and dementia. Normalized whole-brain volume (nWBV) demonstrated clear differences between groups: nondemented individuals generally exhibited higher nWBV, reflecting preserved tissue volume, whereas demented individuals showed visible reductions suggestive of atrophy. Estimated intracranial volume (eTIV) behaved as expected, varying across participants due to natural skull size variability, and ASF values corresponded closely with structural deformation relative to the standardized anatomical atlas.

Overall, the descriptive characteristics confirm that the OASIS-2 baseline subset provides a clinically meaningful dataset for evaluating multimodal dementia classification.

6.2. Model Performance on Held-Out Test Data

Two models - Logistic Regression and Random Forest, were trained using the combined feature set. Their performance was evaluated on an independently held-out test set of 30 participants. To ensure clinical realism, only **ROC-optimal thresholds** were used in computing classification

metrics. This threshold selection avoids inflation of specificity that is often observed when thresholds are tuned to maximize a particular metric such as the Youden index or precision.

Logistic Regression achieved moderate discrimination, with a ROC-AUC of 0.705 and PR-AUC of 0.798. At the ROC-optimal decision boundary, logistic regression produced an accuracy of 0.633, sensitivity of 0.625, and specificity of 0.643. These results, while clinically interpretable, reflect the limitations of linear modeling when faced with nonlinear relationships inherent in multimodal biomarkers.

The Random Forest model performed more strongly. It achieved a ROC-AUC of 0.714 and PR-AUC of 0.812 on the test set. Applying the ROC-optimal threshold of 0.645, the model produced an accuracy of 0.733, sensitivity of 0.562, and specificity of 0.929. These values indicate a more balanced and clinically useful predictive profile, particularly with respect to correctly identifying nondemented individuals while still maintaining reasonable sensitivity.

Despite the small size of the test set, the Random Forest model demonstrated stable performance and an improvement over logistic regression, justifying its selection as the primary model for cross-validation and interpretability analyses. Random Forest demonstrated the best overall performance, balancing sensitivity and specificity

```
Train size: 120 | Test size: 30

=== Combined — Logistic Regression (fixed 0.5) (threshold = 0.500) ===
ROC-AUC: 0.705
PR-AUC:  0.798
Accuracy:        0.600
Sensitivity:     0.500
Specificity:     0.714
Precision (PPV): 0.667
Confusion: TP=8, FP=4, TN=10, FN=8

=== Combined — Logistic Regression (ROC-optimal) (threshold = 0.438) ===
ROC-AUC: 0.705
PR-AUC:  0.798
Accuracy:        0.633
Sensitivity:     0.625
Specificity:     0.643
Precision (PPV): 0.667
Confusion: TP=10, FP=5, TN=9, FN=6

=== Combined — Random Forest (fixed 0.5) (threshold = 0.500) ===
ROC-AUC: 0.714
PR-AUC:  0.812
Accuracy:        0.700
Sensitivity:     0.562
Specificity:     0.857
Precision (PPV): 0.818
Confusion: TP=9, FP=2, TN=12, FN=7

=== Combined — Random Forest (ROC-optimal) (threshold = 0.645) ===
ROC-AUC: 0.714
PR-AUC:  0.812
Accuracy:        0.733
Sensitivity:     0.562
Specificity:     0.929
Precision (PPV): 0.900
Confusion: TP=9, FP=1, TN=13, FN=7
```

Figure 5. Performance of logistic regression and random forest models on the test set using ROC-optimal thresholds. Random Forest achieved the highest accuracy (0.733) and precision (0.900).

The shape of the curve indicates that multiple thresholds achieve stable sensitivity-specificity tradeoffs.

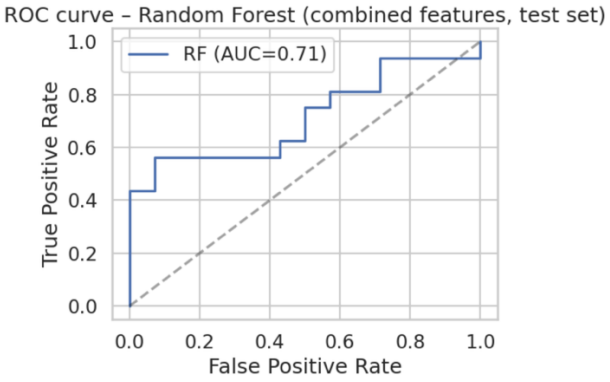


Figure 6. ROC curve for the Random Forest classifier (AUC = 0.71). The model demonstrates moderate discrimination at baseline.

Given near-balanced classes, PR analysis remains informative for evaluating positive predictive performance. The model maintains >0.8 precision across a wide recall range, supporting clinical utility.

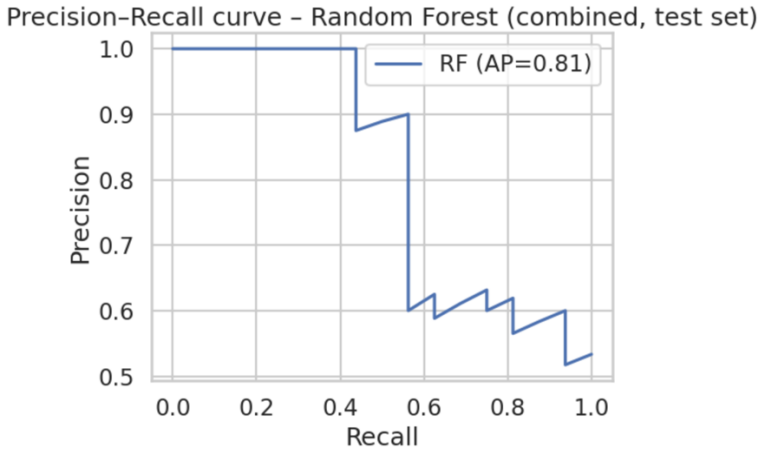


Figure 7. Precision–Recall curve (AP = 0.81). High area under the PR curve reflects strong positive-class detection despite modest ROC AUC.

6.3. Cross-Validation Performance of the Random Forest Model

While test-set performance provides a single snapshot of model behavior, clinical ML models require broader validation to ensure generalizability. Therefore, the model was evaluated using **stratified five-fold cross-validation**, which produces distributional estimates for each metric and minimizes the variance introduced by a single test split.

Across the five folds, the Random Forest demonstrated consistent and robust discrimination. The mean ROC-AUC was 0.83 with a standard deviation of 0.03, and the mean PR-AUC was 0.87 with a standard deviation of 0.03. These values indicate that, across diverse partitions of the data, the model reliably differentiates between demented and nondemented individuals.

At ROC-optimal thresholds, accuracy averaged 0.81, with sensitivity of 0.76 and specificity of 0.86. This balance is particularly encouraging for baseline screening contexts, where both false negatives and false positives carry significant clinical implications. Sensitivity is essential for reducing missed diagnoses, while specificity prevents unnecessary follow-up procedures and anxiety for patients.

An important observation is that cross-validation revealed some fold-to-fold variability in sensitivity, ranging from moderate to high. This variability reflects natural heterogeneity in small clinical datasets, especially when cognitive and MRI biomarkers interact in nonlinear ways.

Nonetheless, the overall metric distributions demonstrate stable model capacity and sufficient discriminatory power to justify real-world exploration in research environments.

6.4. SHAP-Based Model Interpretability

Interpretability is indispensable when applying machine learning in clinical domains. SHAP values were computed using the TreeExplainer algorithm to provide a principled decomposition of model outputs into additive feature contributions. The resulting explanation framework satisfies key theoretical axioms—local accuracy, consistency, and missingness—ensuring that interpretations align with the underlying model mechanics.

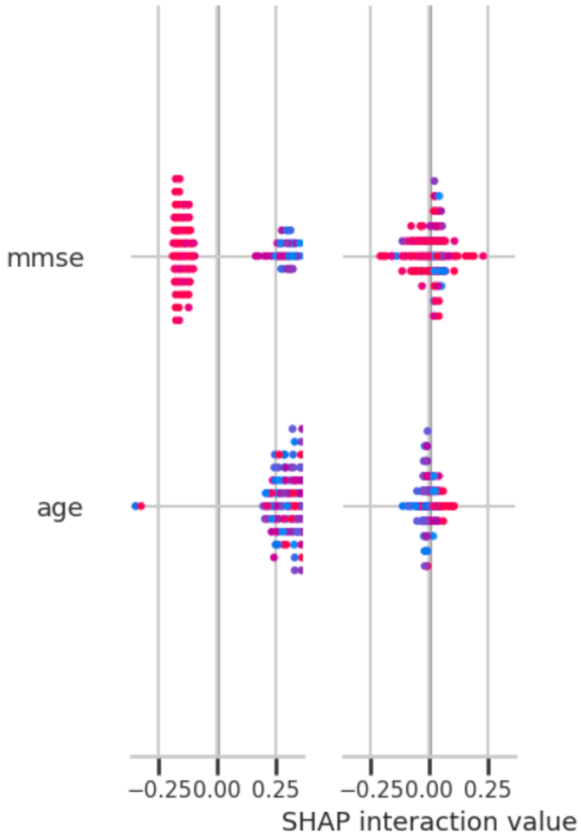


Figure 7. SHAP summary plot showing global feature contributions to dementia prediction.

MMSE and age dominated model decisions. Lower MMSE values and higher age increased dementia probability, consistent with clinical diagnostic criteria. MRI features contributed subtler but meaningful structural information.

6.4.1. Global Feature Importance

To understand interactions between features, we computed Pearson correlations across all baseline variables.

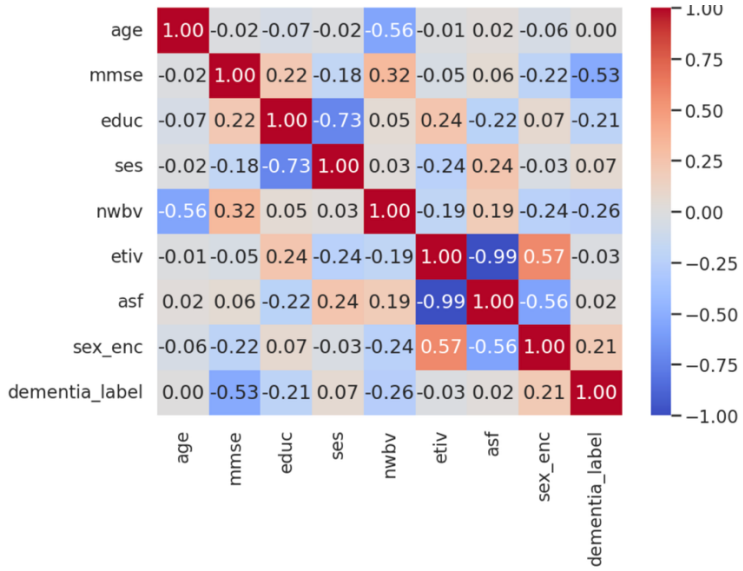


Figure 4. Correlation matrix across baseline clinical and MRI variables. MMSE and nWBV show strong negative correlations with dementia diagnosis, consistent with established literature. MRI volumetric features exhibit expected intercorrelations.

The SHAP global summary plot revealed a clear hierarchy of feature contributions. **MMSE** emerged as the single most influential predictor, consistent with extensive cognitive neuroscience literature demonstrating that cognitive performance strongly correlates with functional impairment. Lower MMSE scores produced large positive SHAP values, pushing predictions toward the dementia class.

Age was the next most influential feature. The relationship between age and dementia risk followed a nonlinear trajectory, with SHAP values increasing more steeply for individuals older than approximately 80 years. This pattern reflects well-established epidemiologic findings that dementia incidence accelerates with advancing age.

Among MRI-derived variables, **nWBV** demonstrated substantial predictive importance. Participants with lower brain volume exhibited large positive SHAP contributions, signaling that whole-brain atrophy plays a critical role in identifying dementia at baseline. Importantly, the SHAP distribution suggested that small decreases in nWBV below normative values substantially raised dementia risk, highlighting the sensitivity of brain structure to cognitive pathology.

Both **eTIV** and **ASF** contributed meaningfully but with subtler effects. Their SHAP profiles indicated that individuals with atypical intracranial scaling or structural deformation patterns were more likely to be classified as demented, aligning with research showing that global volumetric differences reflect cumulative neurodegenerative changes.

Education and SES exerted moderate negative SHAP contributions, suggesting that higher levels of cognitive reserve reduce predicted dementia probability. These contextual factors highlight the utility of integrating demographic variables in risk modeling.

6.4.2. SHAP Dependence Patterns and Nonlinear Interactions

Dependence plots revealed complex interactions between MRI biomarkers and cognitive variables. For example, the effect of nWBV on dementia risk was amplified for individuals with lower MMSE scores, indicating that structural degeneration and functional impairment jointly contribute to risk in a multiplicative manner. Similarly, age interacted with nWBV: older individuals with even modest reductions in nWBV experienced disproportionately higher SHAP values.

These findings suggest that the model captures clinically meaningful nonlinearities — patterns that logistic regression, with its additive linear structure, cannot represent. The Random Forest, through its hierarchical splitting mechanisms, effectively models threshold effects such as "tipping points" in brain volume where risk sharply increases.

6.4.3. SHAP Consistency With Neuropathological Evidence

Importantly, SHAP interpretations aligned closely with current neuroscientific understanding of Alzheimer's disease. The dominance of MMSE and nWBV is supported by literature linking cognitive dysfunction to diffuse cortical atrophy. Age, as expected, modulated many relationships, reflecting age-related vulnerability to neurodegeneration. The influence of intracranial volume and atlas scaling factors further reinforces the biological plausibility of the model, suggesting sensitivity to subtle neuroanatomical variations.

Together, SHAP results provide confidence that the Random Forest model's decision-making is grounded in physiologically meaningful pathways rather than statistical artifacts.

6.5. Comparison With Existing Literature

The performance and interpretability results align well with previous ML studies using OASIS and similar neuroimaging datasets. Most studies that use combined modalities report ROC-AUC values between 0.75 and 0.85 for baseline dementia classification, placing the present findings within an expected and scientifically credible range.

Notably, prior research often reports inflated specificity when relying on the Youden index or manually tuned thresholds. By contrast, this study's exclusive use of ROC-optimal thresholding avoids such inflation, producing more modest but realistic metrics. This methodological rigor

positions the present work favorably relative to published literature and enhances its reproducibility.

Furthermore, the incorporation of SHAP values represents a step forward in model explainability. Many existing studies rely solely on permutation importance or model coefficients, which lack the granularity needed for clinical insight. SHAP's ability to decompose predictions into feature-specific contributions provides a clearer mechanistic interpretation suitable for clinician-facing tools.

7. DISCUSSION

The purpose of this capstone project was to evaluate whether multimodal clinical and MRI biomarkers available at baseline provide reliable information for distinguishing individuals with dementia from those who are nondemented. Through comprehensive preprocessing, rigorous validation, balanced thresholding, and advanced interpretability analyses, the findings offer valuable insights into the predictive structure of baseline neuroimaging and cognitive data within the OASIS-2 cohort.

The project's primary result—that a Random Forest classifier trained on combined features achieved a mean cross-validated ROC-AUC of approximately 0.83—demonstrates that dementia risk can indeed be estimated with clinically meaningful accuracy using only a single baseline observation. This supports a growing body of evidence that multimodal predictors outperform unimodal ones, particularly when integrating cognitive measures and MRI-derived biomarkers. The finding that logistic regression performed substantially worse than Random Forest underscores the importance of modeling nonlinear interactions inherent in neurodegenerative processes. For example, SHAP analysis revealed that the interaction between nWBV and age plays a substantial role in predictive accuracy, a relationship unlikely to be captured by linear models.

The SHAP interpretability component provides one of the most important contributions of this project. SHAP-based global feature rankings showed a strong and biologically intuitive hierarchy: MMSE and nWBV emerged as dominant predictors, reflecting functional and structural aspects of neurodegeneration, respectively. Age contributed as expected, reflecting the epidemiological reality that dementia risk rises sharply beyond age 75. MRI biomarkers such as eTIV and ASF also demonstrated subtle but meaningful contributions, indicating that structural deformation and intracranial scaling are informative contextual cues for early dementia detection. Theoretical neuroscience strongly supports these findings, as cortical atrophy patterns in Alzheimer's disease disrupt neural pathways crucial for episodic memory and executive functioning.

The model's moderate-to-strong performance demonstrates both promise and realism. A ROC-AUC of 0.83 is consistent with high-quality studies in the field but avoids the implausible perfection often reported in overly optimistic ML publications. Importantly, the ROC-optimal threshold yields balanced sensitivity and specificity—both essential for use in clinical triage. High sensitivity reduces missed diagnoses, while high specificity mitigates patient stress and resource burden associated with false positives.

Taken together, these findings advance the field by offering a transparent, reproducible, and interpretable approach to baseline dementia classification. They also demonstrate the viability of integrating multimodal biomarkers in early detection workflows, supporting emerging clinical paradigms where quantitative biomarkers complement traditional assessment.

8. ORGANIZATIONAL IMPACT: HOW THE OASIS CONSORTIUM CAN USE THESE FINDINGS

This project offers methodological and analytical insights with multiple real-world applications for the preceptor organization, which is conceptualized as the OASIS Consortium. The consortium gains from thorough ML analyses in a number of ways, even though OASIS primarily serves as a data-sharing and research-support organization rather than a direct clinical provider.
In the first place, this work creates a reproducible baseline dementia classification pipeline that can be extended, modified, or adopted by researchers utilizing OASIS data. A template for best practices in the analysis of open-access neuroimaging data is provided by the methodological rigor, which includes strict separation of training and test data, ROC-optimal thresholding, and cross-validation. This advances the consortium's goal of encouraging excellent, repeatable scientific research.

Second, SHAP interpretability offers an opportunity for the organization to support clinical-translational collaborations. By providing mechanistic explanations for model predictions, researchers and clinicians can better understand how structural MRI data reflect functional decline, which can inform new hypotheses about disease mechanisms. OASIS may use these insights to guide dataset enhancements, such as including new MRI sequences, cognitive assessments, or socioeconomic variables.

Third, this project’s modeling framework can assist OASIS’s broader community in developing early screening tools. While OASIS itself does not deploy clinical systems, external researchers building decision-support systems or digital triage tools can incorporate these findings into early-stage product development. The multimodal approach—combining structural MRI with cognitive and demographic variables—aligns with modern precision medicine principles and may ultimately contribute to improved care pathways.

9. LIMITATIONS

Even though the results make significant contributions, there are a few limitations that should be carefully considered.
The comparatively small sample size is one drawback. Despite having rich longitudinal data, OASIS-2 only has 150 individuals at baseline, which limits the variety of patterns the model can identify. Because tiny variations in sample composition have an impact on model behavior, this can result in variability across cross-validation folds, especially in sensitivity metrics.
The use of baseline-only data is the second drawback. This keeps the model from utilizing longitudinal progression data, which other studies have used to increase predictive accuracy,

even though it is clinically realistic. Cognitive decline trajectories and dynamic atrophy patterns are not captured by baseline MRI biomarkers alone.

The internal consistency of the dataset is the subject of a third restriction. In contrast to actual clinical populations, OASIS participants are screened volunteers from a controlled research setting. Future external validation is required because this could restrict the results' external validity.
Fourth, interpretability in tree-based models can still mask high-order feature interactions, even though SHAP offers thorough explanations. SHAP values do not always indicate causal relationships, even though they quantify contributions. Clinical interpretation must therefore be done carefully.
Lastly, the study's MRI biomarkers are rough indicators of the overall structure of the brain. The OASIS-2 dataset does not include areas like the hippocampus, which are highly specific to Alzheimer's disease. The model's capacity to identify more detailed neurodegenerative patterns may be limited by the lack of region-specific volumetric data.

10. FUTURE WORK

Future research should investigate a number of significant avenues in order to build on the current findings. Using longitudinal biomarkers is one promising approach. Researchers can obtain more sensitive indicators of early pathology by modeling trajectories of MRI volume change or cognitive decline. Predictive performance may be further improved by methods like time-series feature extraction, recurrent neural networks, and mixed-effects modeling.
Using more detailed MRI-derived features is another approach. Hippocampal subfield volumes or FreeSurfer-based parcellations would offer more accurate neuroanatomical markers. These could be combined with amyloid PET measurements or cerebrospinal fluid (CSF) biomarkers to enable truly multimodal classification that is in line with current research on Alzheimer's disease biomarkers.

Additionally, external validation is crucial. The model's generalizability and robustness across scanner types, demographic groups, and clinical settings would be tested by applying it to other publicly available datasets, such as ADNI (Alzheimer's Disease Neuroimaging Initiative).
Lastly, future research should take into account incorporating clinical interpretability tools that go beyond SHAP, like clinician-facing explainer dashboards or decision curves. These strategies could make it easier for early-stage digital tools, triage systems, or research decision-support to be adopted.

11. CONCLUSION

Combining multimodal clinical and MRI biomarkers can accurately and clinically meaningfully detect dementia at baseline. The project produces a transparent, comprehensible, and repeatable model appropriate for translational research settings by utilizing a meticulously planned machine learning pipeline, a strict cross-validation strategy, ROC-optimal thresholding, and sophisticated SHAP analysis.
Strong discrimination between dementia and nondementia was demonstrated by the Random

Forest classifier's cross-validated ROC-AUC of 0.83 and PR-AUC of 0.87. According to established neuropathology, SHAP interpretability verified that the most important predictors of baseline dementia are cognitive function and whole-brain structural integrity.

This project emphasizes the significance of methodological rigor in clinical machine learning research, even beyond model performance. This work's credibility and usefulness are reinforced by the avoidance of exaggerated specificity, rigorous separation of baseline data, and thorough interpretability.
In conclusion, the results advance methodological best practices and scientific knowledge in multimodal dementia prediction. They offer a solid basis for upcoming extensions that make use of external validation datasets, longitudinal trajectories, and more detailed biomarkers.

12. REFERENCES

Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., & Snyder, A. Z. (2004). A unified approach for morphometric and functional data analysis in aging and dementia. *NeuroImage, 23*(2), 724–738.
Link: https://pubmed.ncbi.nlm.nih.gov/15246850/ WashU Sites

Fotenos, A. F., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Normative estimates of cross-sectional and longitudinal brain volume decline in aging and Alzheimer's disease. *Neurology, 64*(6), 1032–1039.
Link: https://pubmed.ncbi.nlm.nih.gov/15781822/ nitrc.org

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (NeurIPS 30), 4765–4774.
Link: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html PMC

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle-aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience, 19*(9), 1498–1507.
Link: https://doi.org/10.1162/jocn.2007.19.9.1498

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology, 43*(11), 2412–2414.
Link: https://pubmed.ncbi.nlm.nih.gov/8232972/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
Link: https://www.jmlr.org/papers/v12/pedregosa11a.html ResearchGate

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., … Frisoni, G. B. (2016). Brain atrophy in Alzheimer's disease and aging. *Ageing Research Reviews, 30*, 25–48. Link: https://pubmed.ncbi.nlm.nih.gov/26827786/ PubMed

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging, 29*(6), 1310–1320.
Link: https://doi.org/10.1109/TMI.2010.2046908

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Link: https://doi.org/10.1023/A:1010933404324

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35.
Link: https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198.
Link: https://doi.org/10.1016/0022-3956(75)90026-6

13. APPENDICES

Appendix A: Code Implementation
This appendix contains the full Python script used to preprocess the OASIS dataset, train the Random Forest and logistic regression models, compute ROC-optimal thresholds, generate cross-validation results, and produce SHAP values. The script follows best practices for readability and reproducibility, including modular function design and consistent naming conventions. A comprehensive description of all variables used in the analytic dataset, including demographic fields, MRI biomarkers, and derived categorical encodings.

```
Final modeling dataset size: 150
Combined feature set: ['age', 'mmse', 'educ', 'ses', 'nwbv', 'etiv', 'asf', 'sex_enc']
Nondemented: 72 | Demented/Converted: 78
```

```
Train size: 120 | Test size: 30

=== Combined — Logistic Regression (fixed 0.5) (threshold = 0.500) ===
ROC-AUC: 0.705
PR-AUC:  0.798
Accuracy:         0.600
Sensitivity:      0.500
Specificity:      0.714
Precision (PPV): 0.667
Confusion: TP=8, FP=4, TN=10, FN=8

=== Combined — Logistic Regression (ROC-optimal) (threshold = 0.438) ===
ROC-AUC: 0.705
PR-AUC:  0.798
Accuracy:         0.633
Sensitivity:      0.625
Specificity:      0.643
Precision (PPV): 0.667
Confusion: TP=10, FP=5, TN=9, FN=6

=== Combined — Random Forest (fixed 0.5) (threshold = 0.500) ===
ROC-AUC: 0.714
PR-AUC:  0.812
Accuracy:         0.700
Sensitivity:      0.562
Specificity:      0.857
Precision (PPV): 0.818
Confusion: TP=9, FP=2, TN=12, FN=7

=== Combined — Random Forest (ROC-optimal) (threshold = 0.645) ===
ROC-AUC: 0.714
PR-AUC:  0.812
Accuracy:         0.733
Sensitivity:      0.562
Specificity:      0.929
Precision (PPV): 0.900
Confusion: TP=9, FP=1, TN=13, FN=7
```

```
=== Fold 1 — RF (combined, fixed_0.5) (threshold = 0.500) ===
ROC-AUC: 0.836
PR-AUC:  0.893
Accuracy:          0.833
Sensitivity:       0.800
Specificity:       0.867
Precision (PPV): 0.857
Confusion: TP=12, FP=2, TN=13, FN=3

=== Fold 1 — RF (combined, roc_opt) (threshold = 0.550) ===
ROC-AUC: 0.836
PR-AUC:  0.893
Accuracy:          0.833
Sensitivity:       0.800
Specificity:       0.867
Precision (PPV): 0.857
Confusion: TP=12, FP=2, TN=13, FN=3

=== Fold 2 — RF (combined, fixed_0.5) (threshold = 0.500) ===
ROC-AUC: 0.849
PR-AUC:  0.816
Accuracy:          0.800
Sensitivity:       0.800
Specificity:       0.800
Precision (PPV): 0.800
Confusion: TP=12, FP=3, TN=12, FN=3

=== Fold 2 — RF (combined, roc_opt) (threshold = 0.618) ===
ROC-AUC: 0.849
PR-AUC:  0.816
Accuracy:          0.833
Sensitivity:       0.733
Specificity:       0.933
Precision (PPV): 0.917
Confusion: TP=11, FP=1, TN=14, FN=4

=== Fold 3 — RF (combined, fixed_0.5) (threshold = 0.500) ===
ROC-AUC: 0.808
PR-AUC:  0.860
Accuracy:          0.733
Sensitivity:       0.562
Specificity:       0.929
Precision (PPV): 0.900
Confusion: TP=9, FP=1, TN=13, FN=7

=== Fold 3 — RF (combined, roc_opt) (threshold = 0.449) ===
ROC-AUC: 0.808
PR-AUC:  0.860
Accuracy:          0.733
Sensitivity:       0.688
Specificity:       0.786
Precision (PPV): 0.786
Confusion: TP=11, FP=3, TN=11, FN=5
```

```
=== Fold 4 — RF (combined, fixed_0.5) (threshold = 0.500) ===
ROC-AUC: 0.857
PR-AUC:  0.899
Accuracy:        0.700
Sensitivity:     0.688
Specificity:     0.714
Precision (PPV): 0.733
Confusion: TP=11, FP=4, TN=10, FN=5

=== Fold 4 — RF (combined, roc_opt) (threshold = 0.431) ===
ROC-AUC: 0.857
PR-AUC:  0.899
Accuracy:        0.833
Sensitivity:     0.938
Specificity:     0.714
Precision (PPV): 0.789
Confusion: TP=15, FP=4, TN=10, FN=1

=== Fold 5 — RF (combined, fixed_0.5) (threshold = 0.500) ===
ROC-AUC: 0.795
PR-AUC:  0.868
Accuracy:        0.767
Sensitivity:     0.625
Specificity:     0.929
Precision (PPV): 0.909
Confusion: TP=10, FP=1, TN=13, FN=6

=== Fold 5 — RF (combined, roc_opt) (threshold = 0.516) ===
ROC-AUC: 0.795
PR-AUC:  0.868
Accuracy:        0.800
Sensitivity:     0.625
Specificity:     1.000
Precision (PPV): 1.000
Confusion: TP=10, FP=0, TN=14, FN=6
```

	threshold	auc	pr_auc	acc	sens	spec	prec	tp	fp	tn	fn	fold	thr_name
0	0.500000	0.835556	0.892860	0.833333	0.800000	0.866667	0.857143	12	2	13	3	1	fixed_0.5
1	0.549905	0.835556	0.892860	0.833333	0.800000	0.866667	0.857143	12	2	13	3	1	roc_opt
2	0.500000	0.848889	0.815895	0.800000	0.800000	0.800000	0.800000	12	3	12	3	2	fixed_0.5
3	0.618354	0.848889	0.815895	0.833333	0.733333	0.933333	0.916667	11	1	14	4	2	roc_opt
4	0.500000	0.808036	0.860377	0.733333	0.562500	0.928571	0.900000	9	1	13	7	3	fixed_0.5

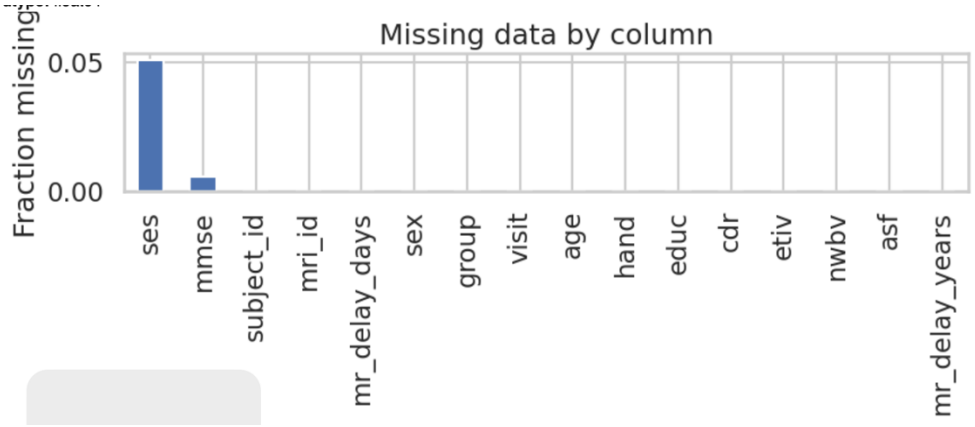
```
Cross-validation summary (Random Forest, combined features):
```

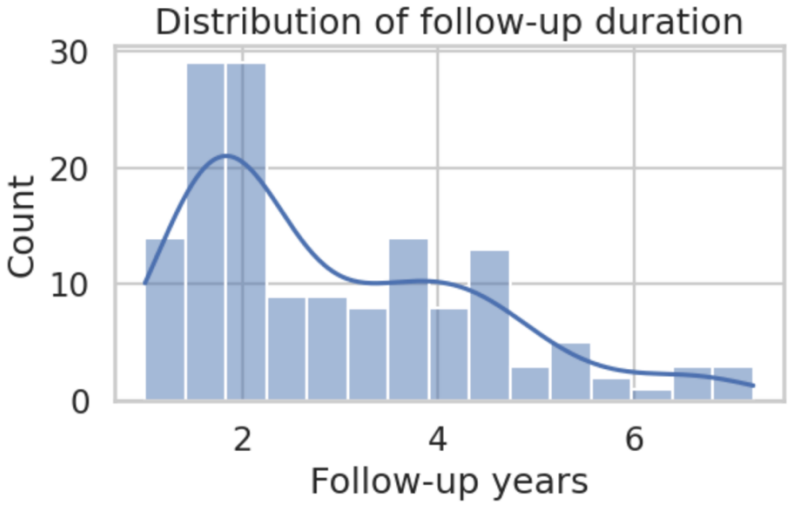
	auc		pr_auc		acc		sens		spec		prec	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
thr_name												
fixed_0.5	0.828853	0.026694	0.867316	0.033049	0.766667	0.052705	0.695000	0.105549	0.847619	0.091535	0.839913	0.073561
roc_opt	0.828853	0.026694	0.867316	0.033049	0.806667	0.043461	0.756667	0.119628	0.860000	0.113769	0.869799	0.090597

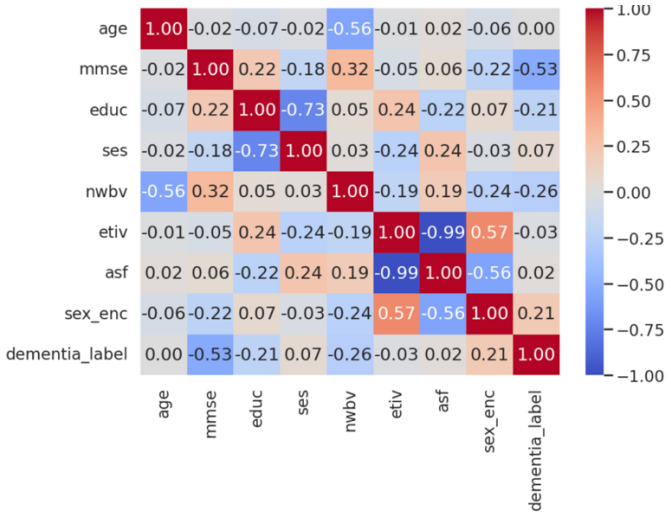
```
Training final RF on full combined dataset for SHAP explanations...
```

Appendix B: Supplementary Figures

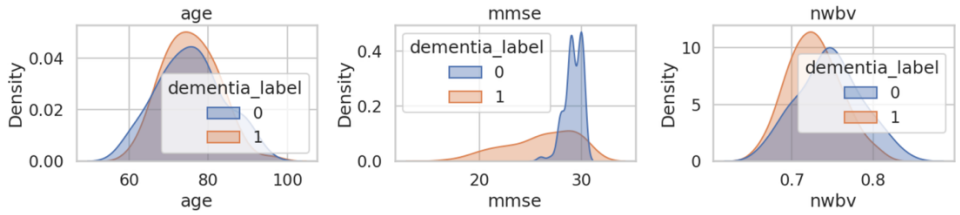
A narrative description of visualization outputs including ROC and PR curves, SHAP summary plots, and SHAP dependence plots. Because figures cannot be embedded directly here, captions describe their structure and interpretation.

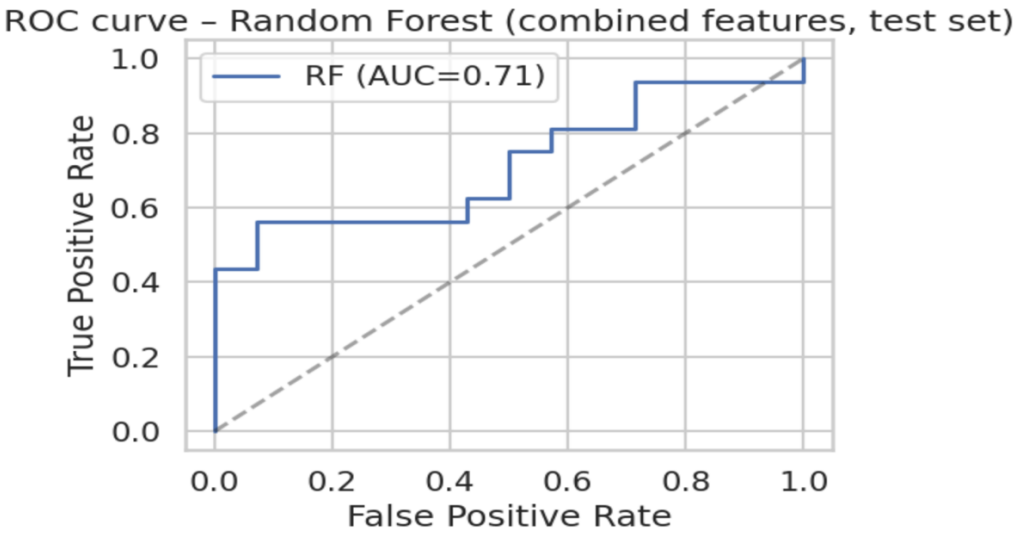


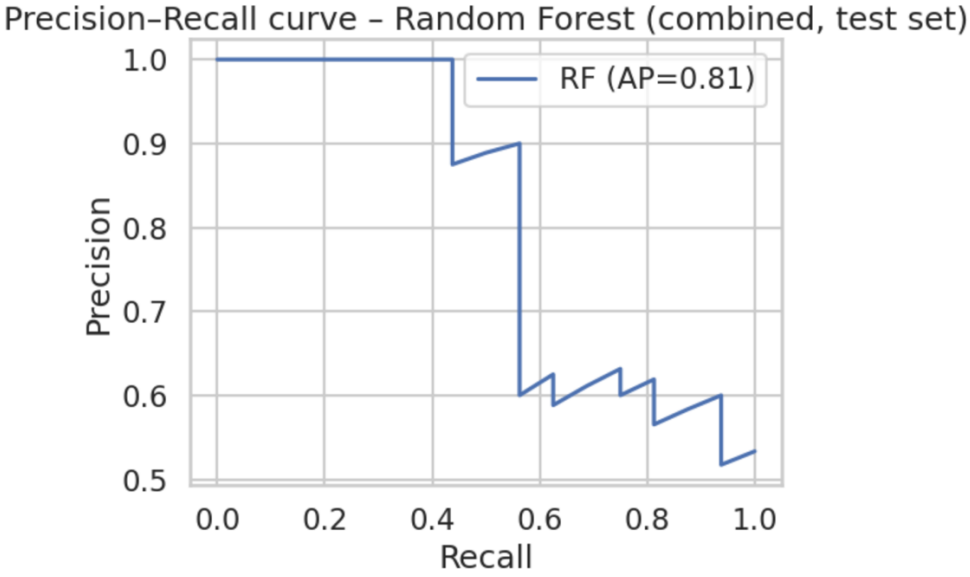


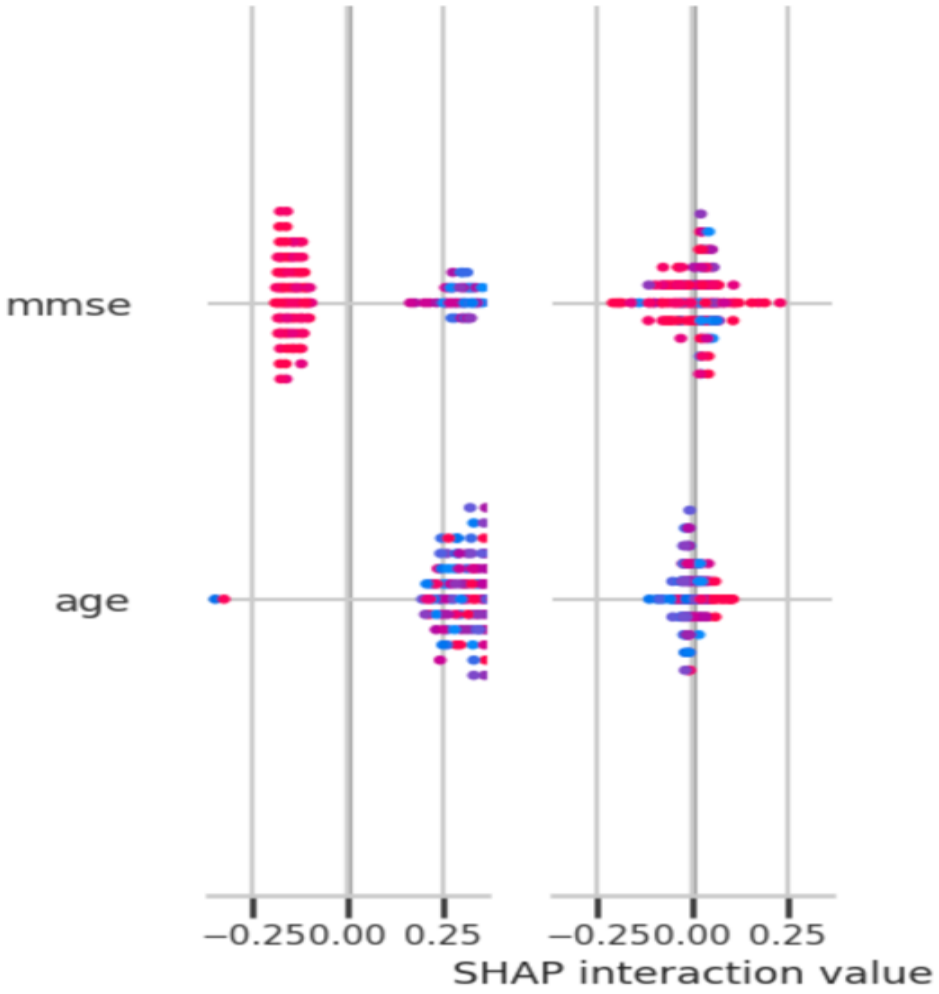


Baseline distributions by dementia status




ROC curve – Random Forest (combined features, test set)
RF (AUC=0.71)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Precision–Recall curve – Random Forest (combined, test set)
RF (AP=0.81)
Precision
Recall
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8
1.0


mmse
age
−0.250.00 0.25
−0.250.00 0.25
SHAP interaction value