



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



MS Health Data Science, Saint Louis University

Vudem Shruthi Reddy ( Msc. Health Data Science )

Instructors: Divya S. Subramaniam, PhD, MPH & Dipti P. Subramaniam, PhD, MPH

Saint Louis University School of Medicine , Department of Health & Clinical outcomes Research

## Background

- **Alzheimer's disease (AD)** is the most common cause of dementia and produces measurable changes in cognition and brain structure several years before clinical diagnosis.
- Early identification of “at-risk” individuals enables timely intervention, improved care planning, and better monitoring of disease progression.
- However, combining clinical + MRI features in a unified predictive model remains limited in real-world datasets because of small sample sizes, collinearity, and class imbalance.
- The OASIS-2 dataset provides longitudinal MRI and cognitive assessments, enabling evaluation of prediction models to classify individuals as demented vs. nondemented at baseline.

## Goals of Study

- Characterize baseline differences in age, cognition, and brain volume between demented and nondemented individuals.
- Develop and evaluate machine-learning models (Logistic Regression, Random Forest) for baseline dementia classification.
- Compare fixed vs. ROC-optimized thresholds for realistic model performance.
- Identify key biomarkers using SHAP explainability to support early risk stratification.

## Analysis

### Statistical & ML Analysis

- Exploratory analysis: group distributions, missing data, correlations.
- **Models:**
  - Logistic Regression (baseline linear model)
  - Random Forest (nonlinear, interaction-aware)
- Evaluation: 80/20 split, ROC-AUC, PR-AUC, accuracy, sensitivity, specificity, PPV.
- **Thresholds:**
  - Fixed 0.50
  - ROC-optimal (Youden Index)
- Cross-validation: 5-fold CV on Random Forest.
- Explainability: SHAP summary & interaction plots.

## Methods

01

### Data Source

- OASIS-2 Longitudinal Dataset (Marcus et al., 2007): 373 MRI sessions from 150 older adults followed up to 7.2 years.
- Data used: baseline visit only (N = 150)

02

### Outcome

- Dementia status at baseline:
  - 0 = Nondemented
  - 1 = Demented / Converted

03

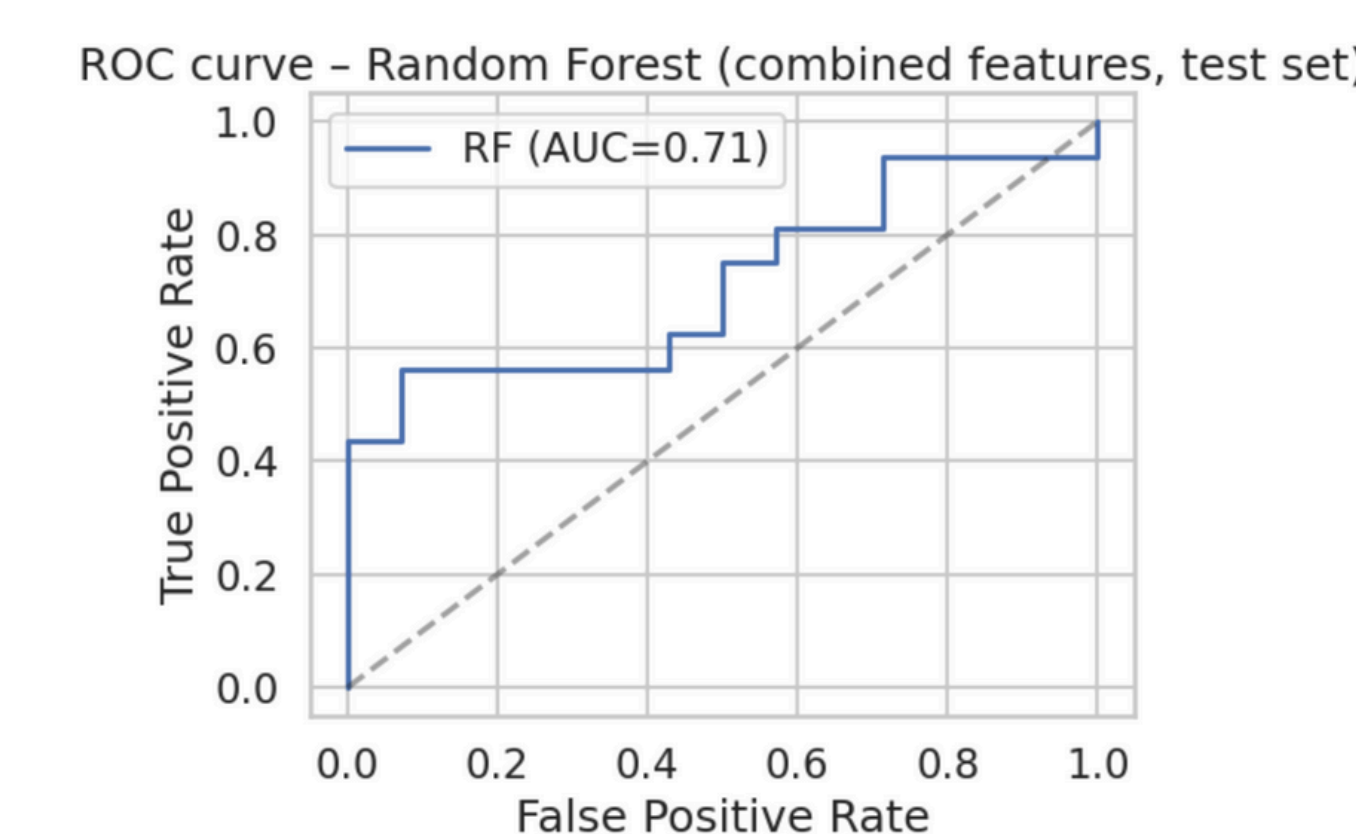
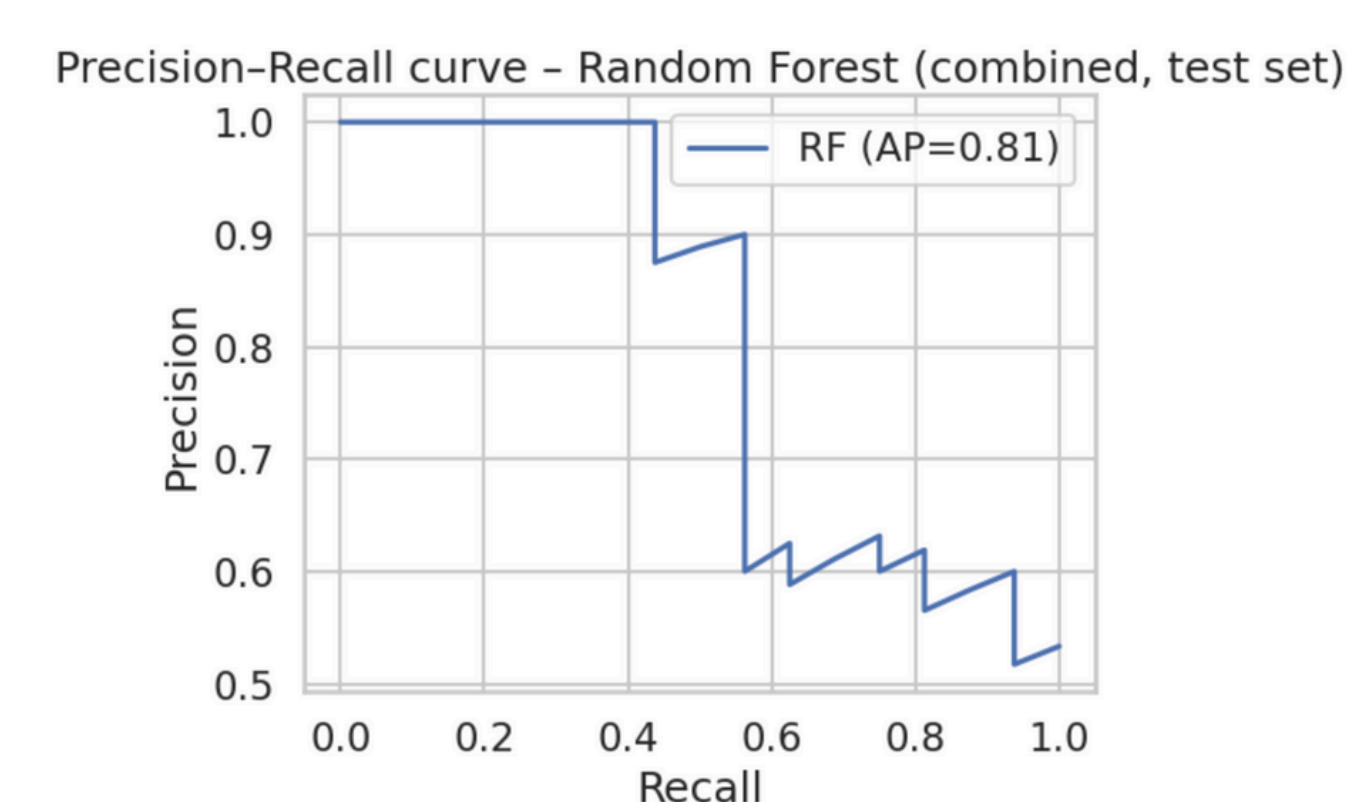
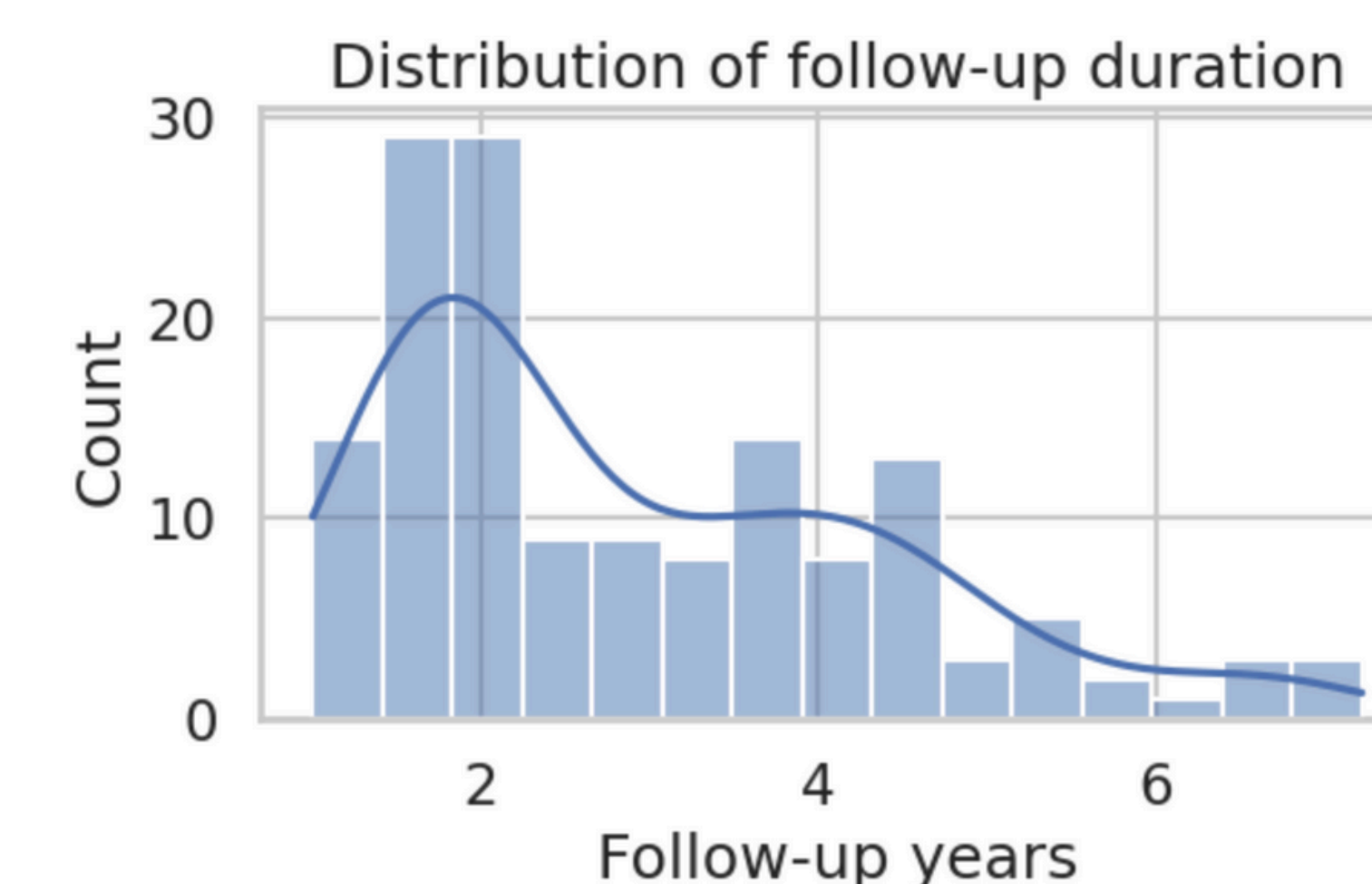
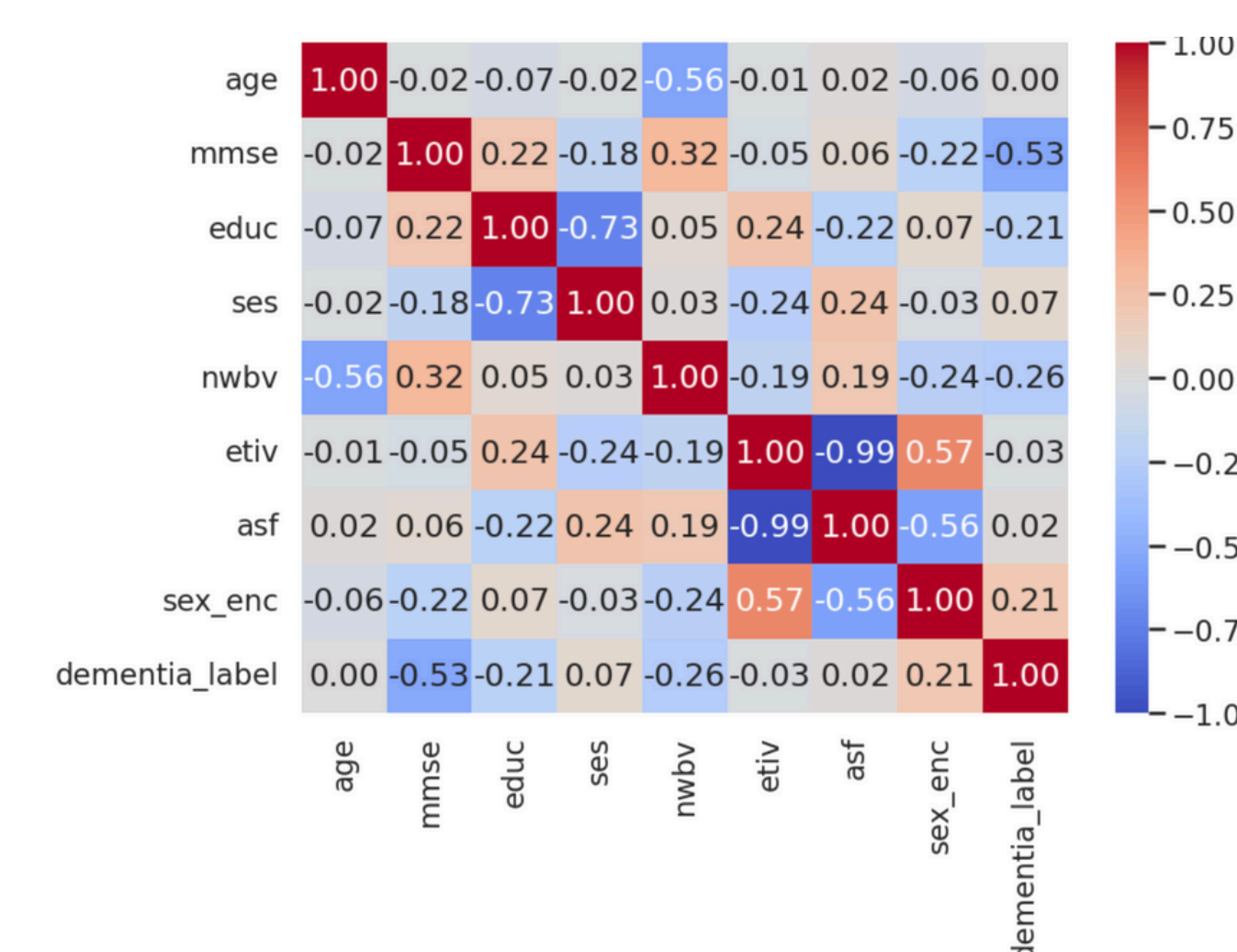
### Predictors

- Clinical:
- Age, MMSE, SES, Education, Sex

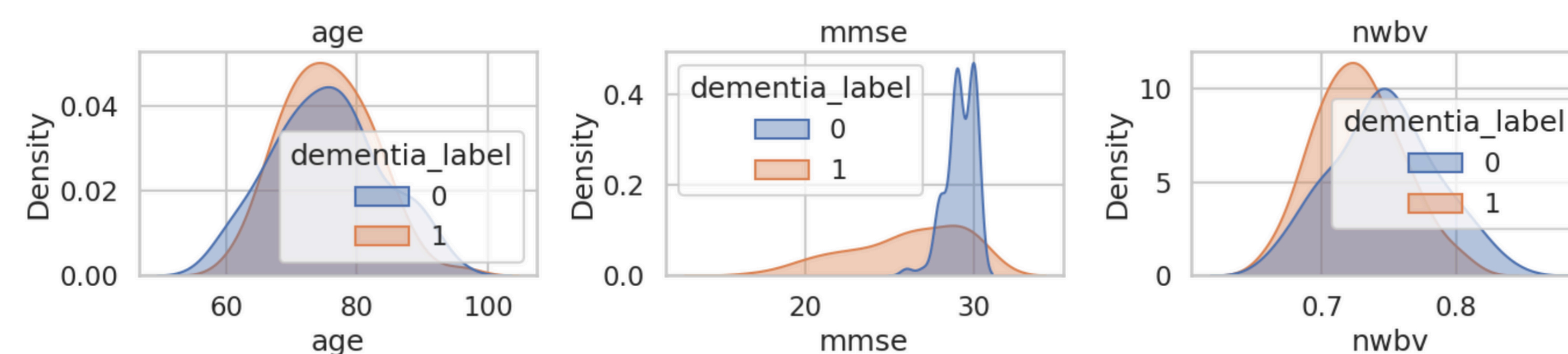
04

### MRI-based:

- Normalized whole-brain volume (nWBV)
- Estimated total intracranial volume (eTIV)
- Atlas scaling factor (ASF)



Baseline distributions by dementia status



## Results

- Demented individuals showed lower MMSE scores, lower nWBV, and higher age, consistent with characteristic Alzheimer's patterns.
- **Logistic Regression** achieved moderate discrimination (ROC-AUC = 0.705).
- **Random Forest** performed best, capturing nonlinear biomarker interactions
- **Test ROC-AUC** = 0.714
- **Test PR-AUC** = 0.812
- **ROC-optimal threshold** improved specificity while maintaining reasonable sensitivity.
- **5-fold cross-validation** demonstrated stable performance (mean ROC-AUC  $\approx$  0.829).
- SHAP analysis identified MMSE, age, and nWBV as the most important predictors of dementia status.

## Interpretation of Findings

- Dementia cases showed lower MMSE scores and reduced normalized whole-brain volume, reflecting established neurodegenerative patterns documented in Alzheimer's disease.
- The Random Forest model outperformed Logistic Regression by capturing nonlinear relationships among brain volume, cognition, and demographic factors.
- ROC-optimal thresholds improved model specificity without substantially reducing sensitivity, producing clinically meaningful classification performance.

## Conclusion

- Multimodal clinical and MRI features provide useful signal for identifying dementia at baseline.
- Random Forest outperformed linear models by capturing nonlinear relationships and biomarker interactions.
- SHAP analysis confirmed MMSE, age, and structural brain volume as key contributors to early dementia classification.
- Findings support the value of multimodal biomarkers for early detection but should be validated on larger and more diverse cohorts.