

Comparing Machine Learning Approaches for Longitudinal Prediction of Clinical Dementia Rating from Multimodal MRI

Ali Eslamian^{1,3}, Qiang Cheng^{1,2}, Colleen Pappas³, Christopher E. Bauer³, Neetu Nair³, Subeksha Khanal³ & Brian T. Gold^{3,4}
¹Computer Science · ²Institute for Biomedical Informatics · ³Neuroscience · ⁴Sanders-Brown Center on Aging · University of Kentucky, Lexington, KY



INTRODUCTION

Early identification of individuals at risk for dementia may be improved by modeling longitudinal trajectories of brain structure and cognitive performance. To this end, we leverage two large repositories (SCAN and ADNI) to test how accurately combined neuroimaging and neuropsychological data predict conversion from normal cognitive to dementia.

AIM

Compare three mixed-effects machine learning models for predicting Clinical Dementia Rating (CDR-Global) from structural MRI, white matter measures, and demographics across repeated visits.

METHODS

Study cohort

604 Participants
71.6 Mean age, yrs
260/344 Male / Female
≥ 3 Visits each

SCAN + ADNI participants with baseline MRI and at least two follow-up visits.

Imaging predictors

- Hippocampal volume
- Entorhinal cortical thickness
- White matter hyperintensities (WMH)
- Other regional volumes & cortical thicknesses

Adjusted for age, sex and education as covariates.

Processing

- FreeSurfer regional volumes and cortical thickness (Desikan–Killiany atlas); WMH from standard segmentation.
- Site effects removed with ComBat harmonization; features residualized for age, sex, education and head size (Total intracranial volume).
- All preprocessing was fit on the training set only to prevent leakage.

Analysis

Outcome: CDR-Global (0–2) was modeled with time-by-predictor interactions and subject-specific random effects. Data were split 80/20 by subject; test predictions use fixed effects only. Model comparison metric: **multiclass accuracy**.

ACKNOWLEDGEMENTS

Data from the ADNI and SCAN consortia. Analysis facilities supported by the University of Kentucky and the Sanders-Brown Center on Aging.

MODELS COMPARED

- 1 GPBoost**
Gradient boosting with grouped random effects
- 2 MERF**
Mixed-Effects Random Forest
- 3 LME**
One-vs-rest logistic linear mixed-effects (BinomialBayesMixedGLM)

RESULTS: CDR TRAJECTORIES

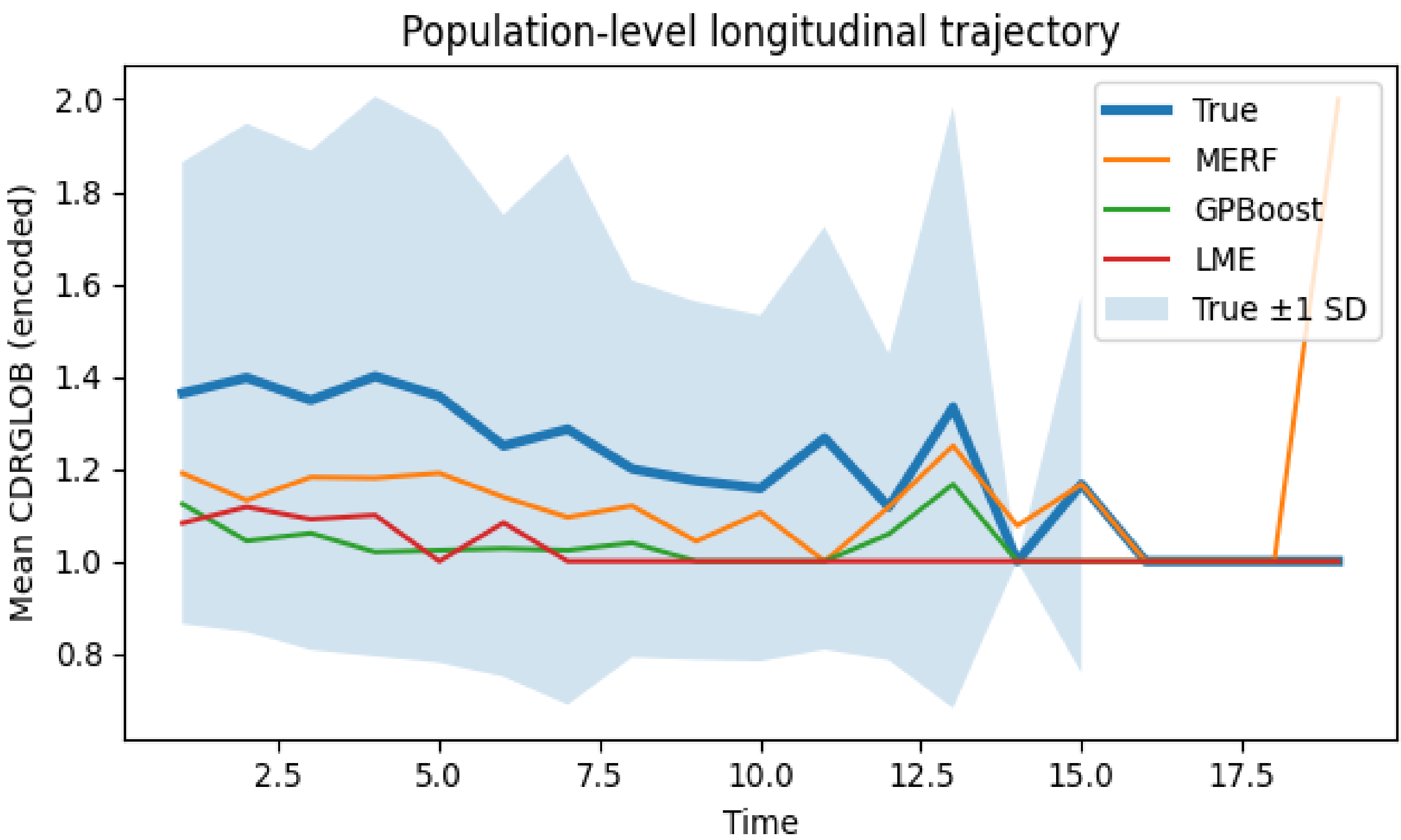


Figure 1. Population-level mean CDR-GLOB trajectories; shaded bands = ± 1 SD. GPBoost and MERF track the rising trend more closely than the flatter LME trajectory.

FEATURE IMPORTANCE

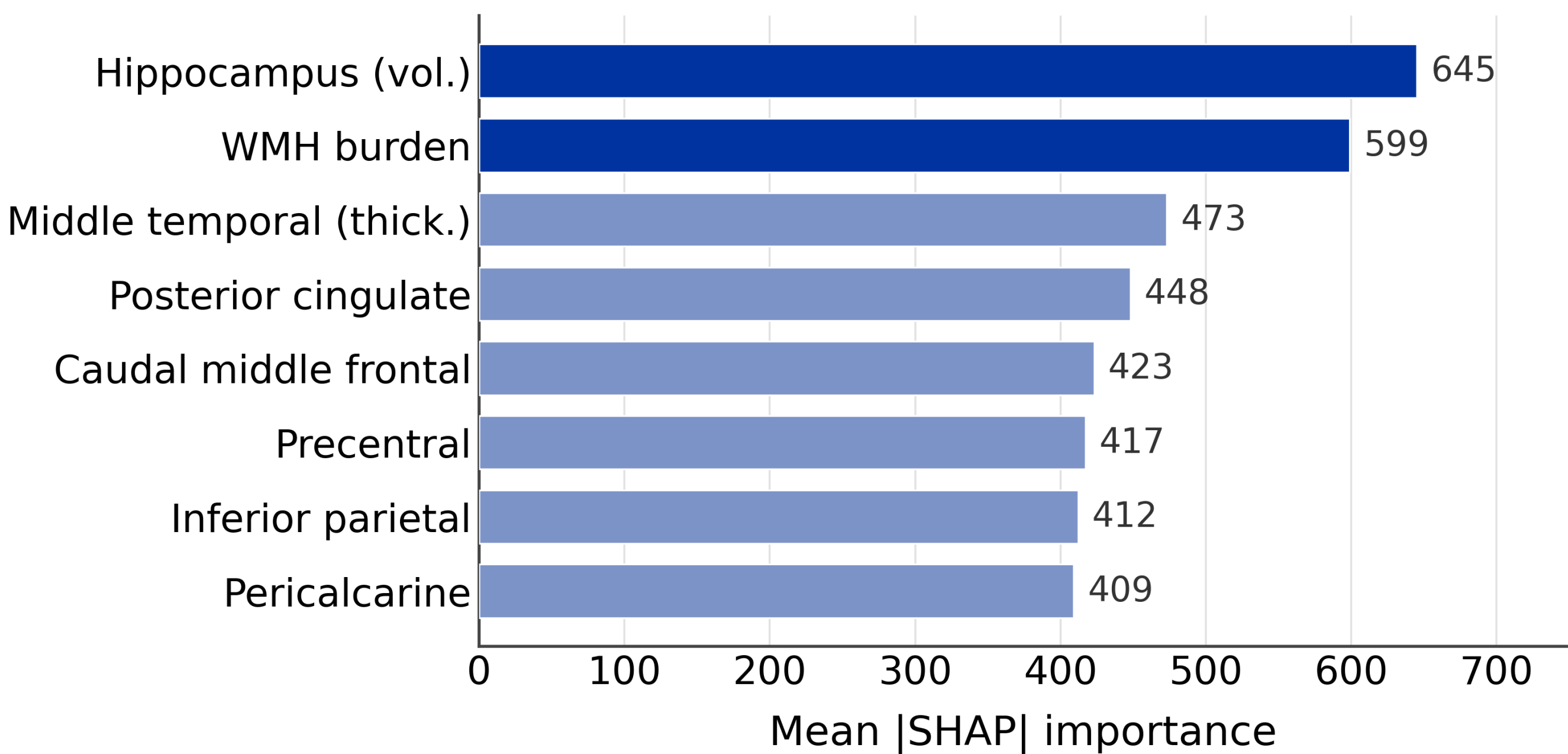


Figure 2. SHAP-derived importance of structural predictors in the GPBoost model. Hippocampal volume and WMH burden (dark) lead, followed by a cluster of cortical-thickness measures (light).

KEY RESULTS

Test-set accuracy

Gaussian Process Boosting (GPBoost) 0.740 ★

Linear Mixed Effects (LME) 0.727

Mixed-Effects Random Forest (MERF) 0.715

Multiclass accuracy on subject-wise test set (range: 0.68 – 0.76)

Findings

- GPBoost gave the highest accuracy (0.740), ahead of LME (0.727) and MERF (0.715).
- Among MRI features, hippocampal volume and WMH burden were the leading predictors (Fig. 2).
- Cortical-thickness measures across temporal, cingulate, frontal and parietal regions formed the next tier.
- Tree-based models captured nonlinear, stage-dependent interactions (medial temporal cortex; a lingual-gyrus hub) that the linear LME cannot represent. (Interactive Fig. 1).
- Data-driven causal analysis placed entorhinal and parahippocampal thickness upstream of hippocampal volume (Interactive Fig. 2).

CONCLUSIONS

- Mixed-effects machine learning integrates longitudinal multimodal MRI with demographics to predict CDR-GLOB.
- GPBoost captures nonlinear, regionally-localized associations and outperforms both LME and MERF.
- A compact set of medial-temporal and cortical-thickness features carries most of the signal, pointing to parsimonious, interpretable models.
- Future work: longer follow-up and additional cohorts.



EXPLORE INTERACTIVE FIGURES

Scan to open the interaction & causal-network figures. Questions and feedback are very welcome.