An Ensemble Machine Learning Model for the Prediction of the Conversion from Mild Cognitive Impairment to Alzheimer's Disease Within 6 Years

Jiayuan Xu[1], Libby J. Walton[1], Andrew J. Doig[1], Sofia Michopoulou[2], Petroula Proitsi[3], Fumie Costen[1]

[1]University of Manchester [2]Medical Physics, University Hospital Southampton NHS Trust & University of Southampton [3]Centre for Preventive Neurology, Wolfson Institute for Population Health, QMUL

Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disease leading to cognitive decline and dementia. Mild Cognitive Impairment (MCI) is an intermediate stage between normal aging and AD. **10–15% of MCI patients convert to AD each year, and ∼80% progress within 6 years.** Early identification is crucial for timely interventions, especially with newly approved disease-modifying treatments (Lecanemab, 2023; Donanemab, 2024). Most prior studies focus on **3-year** prediction. We extend to **6 years** to distinguish long-term stable MCI from progressive MCI, enabling earlier clinical action.

Data & Subjects

Data from the **ADNI** (Alzheimer's Disease Neuroimaging Initiative) ADNIMERGE dataset. **512 MCI patients**: 359 converted to AD within 6 years; 153 remained stable. **35 baseline features** across 8 modalities:

- **Demographic**: age, gender, education, marital status
- **Clinical scales**: CDRSB, FAQ, Everyday Cognition
- **Neuropsychological**: MMSE, ADAS, RAVLT, MoCA, etc.
- **MRI volumes**: hippocampus, ventricles, entorhinal, etc.
- **PET**: FDG-PET metabolism, AV45 amyloid PET
- **CSF**: Aβ42, total-tau, phospho-tau
- **Plasma**: Aβ42/40 ratio, NFL, P-tau181
- **Genetic**: APOE ε4 allele count & baseline diagnosis

Preprocessing

Missing data: Median imputation per variable, addressing multi-modal data sparsity inherent in the ADNI cohort.

Class imbalance: SMOTE oversampling applied *after* train/test split (80%/20%) to prevent data leakage (153 stable vs. 359 converters).

Validation: 5 independent Learning/Testing Sets generated by repeated splitting and over-sampling; final metrics averaged across all 5 Testing Sets.

Encoding & Standardisation: Categorical variables numerically encoded; all features standardised to zero mean and unit variance.

Hyperparameter tuning: Grid search with 10-fold cross-validation on training data; best configuration selected by highest generalised accuracy.

Ensemble Model Architecture

We evaluated **voting** and **stacking** ensembles across all combinations of 5 base classifiers (SVM, LR, KNN, DT, MLP). Best: **stacking** with **LR + SVM + DT + MLP**, achieving accuracy **0.931**.

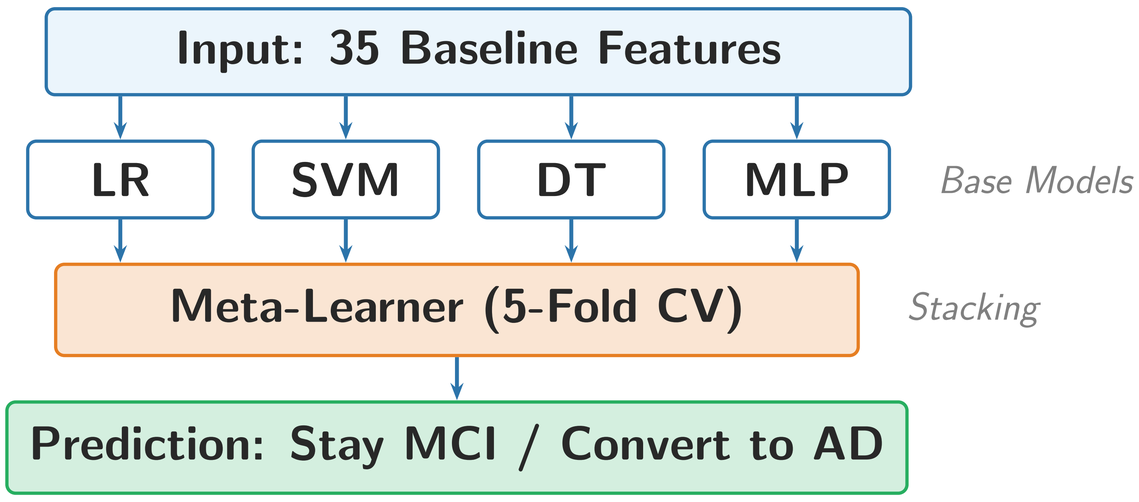


Figure 1. Stacking ensemble model architecture.

Feature Group Analysis

Table 1. Prediction accuracy by feature group (avg. over 5 Testing Sets).

Group	Features Included	Acc.	SD
A	All 35 features	**0.92**	0.02
B	Base (demo + clinical + neuro)	0.82	0.03
C	Base + APOE ε4	0.83	0.04
D	Base + MRI volumes	0.85	0.01
E	Base + PET imaging	0.86	0.03
F	Base + CSF biomarkers	0.85	0.01
G	Base + Plasma biomarkers	0.84	0.02

Comparison with Prior Work

Table 2. Comparison of MCI-to-AD conversion prediction studies.

Study	Yrs	Feature Modalities	Acc.
Li et al. (2021)	3	SNP, mRNA, sMRI	0.792
Gao et al. (2020)	3	Age, sMRI	0.76
Lehallier et al. (2016)	1–6	CSF, plasma	0.80
Chen et al. (2024)	3	MRI, clinical	0.887
Ritter et al. (2015)	3	MRI, PET, CSF, clinical	0.734
Varatharajah et al. (2019)	3	MRI, PET, CSF, clinical	0.93
This work	**6**	**Multimodal (35 features)**	**0.92**

Results

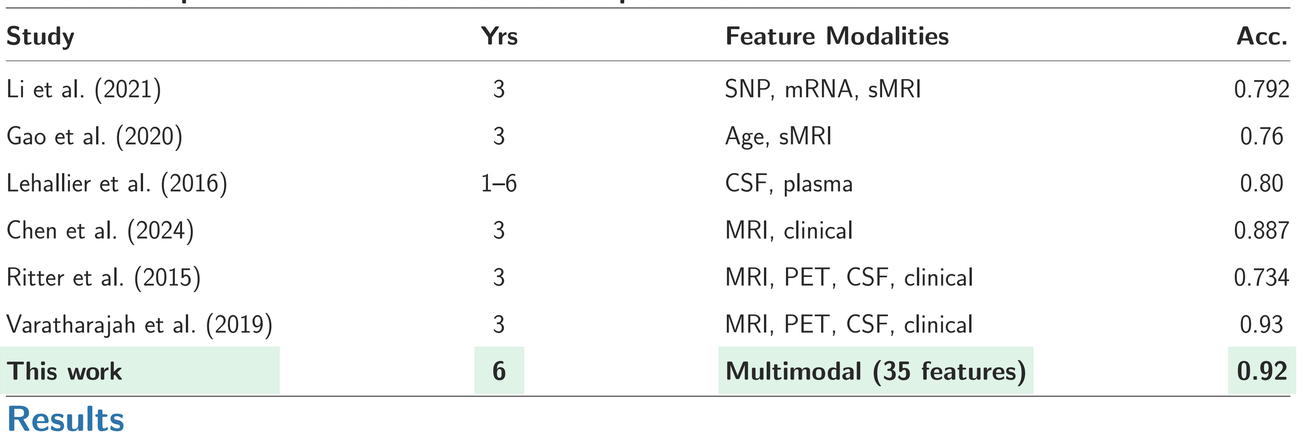
Final Model Performance

(Average over 5 Independent Testing Sets)

Accuracy	Precision	Recall	F1	AUROC
0.92±.02	0.92±.02	0.92±.02	0.91±.01	0.92±.02

Class-level: "Stay MCI" recall = 0.87 (±0.04); "Convert to AD" recall = 0.97 (±0.03). Model is particularly strong at identifying progressive MCI. Best single-run accuracy: **0.94**.

Figure 2. Accuracy comparison across feature groups.



Discussion & Conclusion

- Stacking ensemble achieves **0.92 accuracy for 6-year** MCI-to-AD prediction, outperforming most 3-year models in the literature.
- All 35 multimodal features yield the best result; each modality contributes complementary, non-redundant information.
- Base group alone achieves **0.82** — a cost-effective option for resource-limited clinical settings.
- PET imaging gives the single largest gain (+0.04), highlighting amyloid burden as the strongest individual predictor beyond clinical assessments.

Limitations & Future work: SMOTE was used to address class imbalance; future studies should validate on larger, externally balanced cohorts from ADNI and GAAIN, and explore deep-learning feature integration.

References

[1] Shigemizu et al. *Alzh. Res. Ther.*, 12:145, 2020.
[2] Li et al. *Brain Sci.*, 11(6):674, 2021.
[3] Gao et al. *NeuroImage: Clin.*, 27:102290, 2020.
[4] Chen et al. *Hum. Brain Mapp.*, 45:e26529, 2024.
[5] Varatharajah et al. *Sci. Rep.*, 9:2235, 2019.
[6] Ritter et al. *Alzh. Dem.: DADM*, 1(2), 2015.
[7] Grassi et al. *Front. Neurol.*, 10:756, 2019.
[8] Chawla et al. *JAIR*, 16:321–357, 2002.
[9] Wolpert. *Neural Netw.*, 5(2), 1992.

The University of Manchester

 · jiayuan.xu@postgrad.manchester.ac.uk