

---

# Machine Learning-Based Multiclass Classification of Cognitive Stages Using Plasma Biomarkers, Clinical Assessments, and Genetic Features: A Repeated Nested Cross-Validation Study in ADNI with External Validation in CNTN

---

[Jiayuan Xu](#) and [Fumie Costen](#) \*

Posted Date: 8 May 2026

doi: 10.20944/preprints202605.0498.v1

Keywords: Alzheimer's disease; mild cognitive impairment; plasma biomarkers; machine learning; nested cross-validation; multiclass classification



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

# Machine Learning-Based Multiclass Classification of Cognitive Stages Using Plasma Biomarkers, Clinical Assessments, and Genetic Features: A Repeated Nested Cross-Validation Study in ADNI with External Validation in CNTN

Jiayuan Xu  and Fumie Costen \* 

Department of Electrical and Electronic Engineering, University of Manchester, Manchester, United Kingdom

\* Correspondence: fumie.costen@manchester.ac.uk

## Abstract

**Background:** Plasma biomarkers are widely promoted as scalable tools for staging Alzheimer's disease (AD), yet head-to-head comparisons against the clinical scales used to *define* diagnostic labels remain scarce. Reported gains from machine-learning fusion of clinical and biomarker features may therefore reflect label circularity rather than biological signal. **Methods:** From the Alzheimer's Disease Neuroimaging Initiative (ADNI), we assembled 655 participants (CN = 296, MCI = 168, AD = 191) with concurrent plasma biomarkers (pT217, A $\beta$ 42/40, NfL, GFAP), clinical scales (MMSE, CDR-SB, FAQ), APOE genotype, and demographics. Three pre-specified feature sets (clinical-only, biomarker-plus-demographic-genetic, and full fusion) were compared across four classifiers (Logistic Regression, SVM, Random Forest, XGBoost) using repeated nested cross-validation (5-fold  $\times$  3 outer, 5-fold inner) with balanced class weighting. External validation used the Center for Neurodegeneration and Translational Neuroscience (CNTN) cohort ( $n = 130$ ). **Results:** Clinical scales alone reached a three-class AUC-OvR of  $0.9539 \pm 0.0041$ , and fusion reached  $0.9559 \pm 0.0046$ , an indistinguishable gain. Because MMSE, CDR-SB, and FAQ partly determine ADNI diagnostic labels, both estimates are circularity-inflated upper bounds. Independently of this circularity, the plasma-plus-demographic-genetic model still achieved AUC-OvR =  $0.7455 \pm 0.0150$ , with pT217 the dominant contributor. Pairwise discrimination was excellent for CN vs. AD (1.0000) and MCI vs. AD (0.9739), but markedly weaker for CN vs. MCI (0.9302 fused, 0.69 plasma-only). The reduced biomarker model transferred to CNTN with AUC-OvR = 0.702 (95% CI 0.635–0.764). **Conclusions:** Apparent fusion gains in ADNI are largely a consequence of label circularity. After removing the circular clinical features, plasma pT217 supports three-class CN/MCI/AD screening at AUC  $\approx$  0.74 internally and 0.70 externally, which establishes a realistic performance ceiling for blood-based AD staging. MCI detection remains the principal bottleneck.

**Keywords:** Alzheimer's disease; mild cognitive impairment; plasma biomarkers; machine learning; nested cross-validation; multiclass classification

## 1. Introduction

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder worldwide, affecting an estimated 55 million people globally and projected to exceed 150 million by 2050 [1,2]. The disease follows a protracted continuum progressing from a preclinical phase of normal cognition (CN) through mild cognitive impairment (MCI) to AD dementia [3]. Early and accurate stratification along this continuum is critical for clinical trial enrichment, timely intervention, and the growing era of disease-modifying therapies [4].

Historically, AD staging has relied on costly or invasive procedures such as cerebrospinal fluid (CSF) analysis or amyloid positron emission tomography (PET) imaging [5]. The advent of ultrasensitive immunoassay platforms has enabled detection of AD-related proteins in peripheral blood at clinically actionable concentrations [6]. Among these, phosphorylated tau-217 (pT217), the amyloid- $\beta$  42/40 ratio ( $A\beta_{42/40}$ ), neurofilament light chain (NfL), and glial fibrillary acidic protein (GFAP) have each shown associations with AD pathology [7–9]. However, the extent to which these biomarkers can independently stratify patients across the full CN–MCI–AD spectrum, relative to established cognitive instruments, has not been fully characterized.

Machine learning (ML) models have been widely applied to AD classification tasks using imaging, genetic, and clinical data [10,11]. However, many prior studies employ single train-test splits or simple cross-validation schemes that produce optimistic performance estimates, particularly in small-to-moderate cohorts [12]. Repeated nested cross-validation (CV) mitigates this bias by strictly separating hyperparameter tuning from performance evaluation [13], yet few studies have applied it to plasma biomarker panels for three-class cognitive staging.

The APOE  $\epsilon 4$  allele is the strongest known genetic risk factor for sporadic AD [14], and incorporating it alongside plasma biomarkers and demographic variables may improve classification beyond biomarkers alone. Whether such a composite biomarker model approaches clinical-assessment-level performance has important implications for scalable, minimally invasive cognitive screening.

The present study addresses three pre-specified objectives: (1) to evaluate whether plasma biomarkers augmented by demographic and genetic information can classify CN, MCI, and AD; (2) to compare this biomarker-based model against one relying exclusively on validated clinical scales; and (3) to determine whether a fusion of all feature types yields additional discriminative gain. These objectives are addressed using a repeated nested CV protocol applied to the ADNI cohort, with external evaluation in the CNTN cohort, and 95% confidence intervals are reported for all performance estimates.

## 2. Materials and Methods

### 2.1. Data Sources

#### 2.1.1. ADNI Cohort

The primary analysis cohort was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI; <https://adni.loni.usc.edu>) [15], a longitudinal multi-site study launched in 2003 to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and monitoring of AD. ADNI was approved by the institutional review boards of all participating institutions, and written informed consent was obtained from all participants or their authorized representatives.

ADNI phases included were ADNI 1, ADNI GO, ADNI 2, ADNI 3, and ADNI 4. Diagnosis records were downloaded in February 2026 (DXSUM\_02Feb2026.csv). Plasma biomarker data were obtained from the University of Pennsylvania dataset using Fujirebio Lumipulse and Quanterix Simoa platforms (UPENN\_PLASMA\_FUJIREBIO\_QUANTERIX\_02Feb2026.csv).

#### 2.1.2. CNTN Cohort

To assess robustness and external generalizability of the ADNI-derived models, an independent dataset was obtained from the Center for Neurodegeneration and Translational Neuroscience (CNTN; <https://nevadacntn.org/>) data center [16], dedicated to studying neurodegenerative diseases in the aging population. The CNTN study protocols were approved by the Cleveland Clinic Institutional Review Board, and all participants provided written informed consent.

CNTN measures plasma pTau-181 rather than pTau-217 and does not include the plasma  $A\beta_{42/40}$  ratio, so external validation was performed on the intersection of variables available in both datasets (NfL, GFAP, APOE genotype, age, sex, education). After applying inclusion criteria analogous to ADNI, the analytic external sample comprised  $n = 130$  participants (CN = 57, MCI = 56, AD = 17).

## 2.2. Participants and Diagnostic Labels

Diagnosis codes follow the ADNI convention: 1 = CN, 2 = MCI, 3 = AD. Eligible participants required concurrent availability of a valid diagnosis label, plasma biomarker concentrations (pT217, A $\beta$ 42/40, NfL, GFAP), clinical assessments (MMSE, CDR-SB, FAQ), APOE genotype, and demographic data (age, sex, years of education). ADNI encodes missing or unavailable biomarker values as sentinel codes ( $-4.0$  = not available,  $-5.0$  = not done). All rows containing sentinel values in any plasma biomarker column were excluded prior to analysis. Plasma samples and diagnostic assessments were obtained at the same ADNI visit (matched by visit code VISCODE2). A total of 655 unique participants met all inclusion criteria: CN = 296, MCI = 168, AD = 191.

## 2.3. Feature Variables

Three feature sets were pre-specified to enable systematic ablation:

*Clinical-only* (3 features): MMSE score, CDR-SB, and FAQ total score. These instruments constitute the backbone of cognitive staging in clinical practice and ADNI study criteria.

*Biomarker-plus-demographic-genetic* (8 features): plasma pT217 (pg/mL), A $\beta$ 42/40 ratio, NfL (pg/mL), GFAP (pg/mL), APOE4 allele count (0, 1, or 2), age at examination (years), years of education, and sex (1 = male, 0 = female).

*Fusion* (11 features): the union of all variables from the two sets above.

## 2.4. Statistical Characterization of Participants

Between-group differences for continuous variables were assessed by one-way analysis of variance (ANOVA), with Tukey's honestly significant difference (HSD) test applied post-hoc for plasma biomarkers. Categorical variables (sex, APOE4 carrier status, APOE4 homozygosity) were compared using Pearson's chi-square test. All  $p$ -values are two-sided;  $p < 0.05$  was considered statistically significant.

## 2.5. Machine Learning Classifiers

Four classifiers were evaluated: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). These models were selected to span a range from linear (LR, SVM) to non-linear ensemble architectures (RF, XGBoost), enabling comparison across model complexity levels [11].

## 2.6. Repeated Nested Cross-Validation Protocol

Model evaluation used a repeated nested CV protocol [12,13], which strictly separates hyperparameter tuning from performance evaluation to avoid optimistic bias.

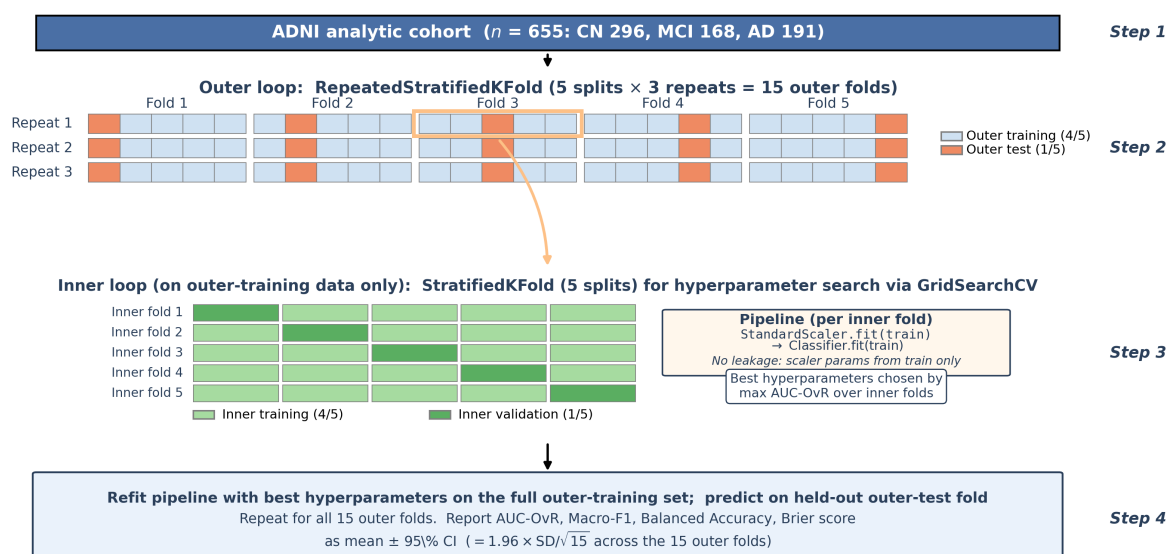
*Outer loop:* RepeatedStratifiedKFold with 5 splits  $\times$  3 repetitions = 15 outer folds.

*Inner loop:* StratifiedKFold with 5 splits applied exclusively to outer training data for hyperparameter tuning via GridSearchCV. Each model was wrapped in a scikit-learn Pipeline consisting of a StandardScaler (z-score normalization) followed by the classifier, ensuring that scaling parameters were fit solely on training data within each fold and applied to the held-out test set, thereby preventing data leakage. For LR, SVM, and RF, class imbalance was handled via `class_weight = 'balanced'`; for XGBoost, per-fold sample weights computed from `sklearn.utils.class_weight.compute_sample_weight` were passed as fit parameters. The two approaches are functionally equivalent.

*Hyperparameter search spaces:* LR searched  $C \in \{0.1, 1, 10\}$  and penalty  $\in \{\ell_1, \ell_2\}$  (SAGA solver). SVM searched  $C \in \{0.5, 1, 5\}$  and kernel  $\in \{\text{linear}, \text{RBF}\}$ . RF searched estimators  $\in \{100, 300\}$ , max depth  $\in \{5, \text{None}\}$ , and min samples split  $\in \{2, 10\}$ . XGBoost searched estimators  $\in \{100, 300\}$ , max depth  $\in \{3, 5\}$ , learning rate  $\in \{0.03, 0.1\}$ , and subsample  $\in \{0.8, 1.0\}$ . Inner-loop optimization used AUC-OVR for three-class tasks and standard AUC for binary tasks.

The full repeated nested protocol is summarised schematically in Figure 1. Three outer-loop repeats (yielding 15 outer folds in total) were used to attenuate the variance contributed by random fold

partitioning: a single 5-fold split is itself a random draw, and pooling 15 fold-level estimates produces a more stable mean and a 95% confidence interval ( $1.96 \times SD/\sqrt{15}$ ) that reflects both within-partition and between-partition variability [12].



**Figure 1.** Schematic of the repeated nested cross-validation protocol. **Step 1:** the full ADNI analytic cohort ( $n = 655$ ). **Step 2:** the outer loop uses RepeatedStratifiedKfold with 5 splits and 3 repeats, producing 15 distinct outer folds; in each fold, four-fifths of the data are retained for training (light blue) and one-fifth is held out as the unseen test set (orange). **Step 3:** within each outer training set, a second StratifiedKfold (5 splits) is applied to perform GridSearchCV-based hyperparameter selection; for every inner fold, a Pipeline fits the StandardScaler and the classifier exclusively on the inner-training portion, preventing any information leakage from the inner-validation split. **Step 4:** the pipeline is refit on the full outer-training set with the best hyperparameters and evaluated on the held-out outer-test fold; this is repeated for all 15 outer folds, and AUC-OvR, Macro-F1, Balanced Accuracy, and the Brier score are reported as mean  $\pm$  95% CI ( $= 1.96 \times SD/\sqrt{15}$ ).

### 2.7. Performance Metrics

On each outer test fold, three discrimination metrics were computed: (1) AUC-OVR (macro-averaged one-vs.-rest area under the ROC curve) for three-class tasks, or standard binary AUC; (2) Macro-F1 score; and (3) Balanced accuracy. Results are reported as mean  $\pm$  95% CI across 15 outer folds, where  $CI = 1.96 \times SD/\sqrt{15}$ . As a calibration metric, (4) the multiclass Brier score was additionally computed for the three-class tasks; binary Brier scores were computed for pairwise tasks. Brier score results are reported in Supplementary Tables S3 (multiclass) and S4 (pairwise).

### 2.8. Pairwise Binary Classification

Pairwise binary classifiers were trained on three tasks (CN vs. MCI, CN vs. AD, MCI vs. AD) using the fusion feature set, applying the same repeated nested CV protocol.

### 2.9. Plasma Biomarker Subset Analysis

To isolate the contribution of each plasma biomarker while controlling for demographic and genetic variables, all 15 non-empty subsets of {pT217, A $\beta$ 42/40, NfL, GFAP} were evaluated using a Random Forest classifier for the three-class (CN/MCI/AD) classification task. In each model, age, sex, education, and APOE4 allele count were included as fixed background features. Random Forest was selected for its strong performance across feature sets and its compatibility with SHAP-based interpretability.

### 2.10. Feature Interpretability via SHAP

SHAP (SHapley Additive exPlanations) values [17] were computed within the same repeated nested cross-validation framework used for performance evaluation. For each of the 15 outer folds,

a Pipeline (StandardScaler + Random Forest with `class_weight = 'balanced'`) was tuned by the same inner-loop 5-fold GridSearchCV used for performance evaluation (§2.6) and refit on the full outer-training partition. TreeSHAP was then applied to the held-out outer-test partition (out-of-fold SHAP), so that each fold's SHAP values come from the same per-fold hyperparameter configuration used for the performance evaluation. For each fold, global feature importance was obtained by averaging absolute SHAP values across both samples and classes, whereas per-class importance was obtained analogously by averaging across samples only (preserving the class axis). Fold-level values were then aggregated across the 15 outer folds and reported as mean  $\pm$  95% CI ( $= 1.96 \times SD/\sqrt{15}$ ). The same procedure was applied to both the biomarker-plus-demographic-genetic and fusion models. Beeswarm visualisations use out-of-fold SHAP averaged across the three test-fold appearances of each participant.

### 2.11. External Validation Procedure

To assess generalisability of the ADNI-derived models, we performed a partial external validation on the CNTN cohort introduced in §2.1.2. Because direct replication of the full primary feature panel was infeasible (CNTN measures pTau-181 rather than pTau-217 and does not include plasma A $\beta$ 42/40), we re-trained the ADNI model on the intersection of available features, namely NfL, GFAP, APOE  $\epsilon$ 4 allele count, age, sex, and education, and applied the resulting pipeline to CNTN without any fine-tuning (zero-shot transfer). For each classifier (LR, SVM, RF, XGBoost) we performed 5-fold inner cross-validation on full ADNI for hyperparameter selection, refit on all of ADNI, and evaluated on CNTN. Multi-class AUC-OvR, Macro-F1, Balanced Accuracy and the multi-class Brier score were computed. The 95% confidence intervals were obtained from 1000 stratified bootstrap resamples of the CNTN evaluation set.

## 3. Results

### 3.1. Participant Characteristics

Table 1 presents demographic, clinical, and biomarker characteristics for the 655 participants by diagnosis group.

**Table 1.** Demographic, clinical, and biomarker characteristics of study participants stratified by diagnosis (one sample per participant).

Variable	CN ( $n = 296$ )	MCI ( $n = 168$ )	AD ( $n = 191$ )	$p$ -value
<i>Demographic and clinical</i>				
Age (years)	77.3 $\pm$ 7.8	79.3 $\pm$ 7.5	79.1 $\pm$ 8.0	0.011
Education (years)	16.7 $\pm$ 2.4	16.0 $\pm$ 2.7	16.1 $\pm$ 2.6	0.008
MMSE score	28.8 $\pm$ 1.5	27.0 $\pm$ 2.7	19.1 $\pm$ 6.3	<0.001
CDR-SB	0.1 $\pm$ 0.3	1.5 $\pm$ 1.2	7.4 $\pm$ 3.5	<0.001
FAQ total	0.4 $\pm$ 1.1	3.9 $\pm$ 5.0	18.9 $\pm$ 7.4	<0.001
<i>Plasma biomarkers</i>				
pT217 (pg/mL)	0.2 $\pm$ 0.2	0.4 $\pm$ 0.3	0.8 $\pm$ 0.6	<0.001
A $\beta$ 42/40 ratio	0.085 $\pm$ 0.013	0.083 $\pm$ 0.011	0.081 $\pm$ 0.012	0.002
NfL (pg/mL)	20.7 $\pm$ 13.6	26.4 $\pm$ 14.5	35.7 $\pm$ 24.5	<0.001
GFAP (pg/mL)	184.2 $\pm$ 102.0	231.0 $\pm$ 140.4	287.8 $\pm$ 175.1	<0.001
<i>Genetic and sex</i>				
Female sex, $n$ (%)	155 (52.4%)	75 (44.6%)	66 (34.6%)	<0.001
APOE4 carrier, $n$ (%)	107 (36.1%)	69 (41.1%)	123 (64.4%)	<0.001
APOE4 homozygous, $n$ (%)	9 (3.0%)	15 (8.9%)	37 (19.4%)	<0.001

Continuous variables: mean  $\pm$  SD; one-way ANOVA  $p$ -value. Categorical variables:  $n$  (%); Pearson chi-square  $p$ -value. Abbreviations: CN, cognitively normal; MCI, mild cognitive impairment; AD, Alzheimer's disease; MMSE, Mini-Mental State Examination; CDR-SB, Clinical Dementia Rating Sum of Boxes; FAQ, Functional Activities Questionnaire; pT217, plasma phosphorylated tau-217; A $\beta$ 42/40, amyloid- $\beta$  42/40 ratio; NfL, neurofilament light chain; GFAP, glial fibrillary acidic protein.

CN participants were younger than MCI and AD participants (77.3 vs. 79.3 and 79.1 years;  $p = 0.011$ ) and had more years of education (16.7 vs. 16.0 and 16.1;  $p = 0.008$ ). All three clinical scales showed a monotonic gradient from CN through MCI to AD (all  $p < 0.001$ ): CDR-SB ranged from 0.1 (CN) to 7.4 (AD) and FAQ from 0.4 (CN) to 18.9 (AD).

Among plasma biomarkers, pT217, NfL, and GFAP differed across groups (all  $p < 0.001$ ). The A $\beta$ 42/40 ratio showed a significant overall group difference ( $p = 0.002$ ), with a subtle decreasing trend from CN to AD consistent with amyloid pathology, although pairwise effect sizes were small.

APOE4 carrier rates were 36.1% (CN), 41.1% (MCI), and 64.4% (AD). APOE4 homozygosity rates were 3.0%, 8.9%, and 19.4%, respectively (both  $p < 0.001$ ). The proportion of female participants was 52.4% (CN), 44.6% (MCI), and 34.6% (AD;  $p < 0.001$ ).

### 3.2. Three-Class Classification Performance

Table 2 presents AUC-OVR, Macro-F1, and Balanced Accuracy from the repeated nested CV experiment across all three feature sets and four classifiers.

**Table 2.** Three-class (CN/MCI/AD) repeated nested cross-validation performance by feature set and classifier ( $n = 655$ ; 15 outer folds).

Feature Set	Classifier	AUC-OVR (95% CI)	Macro-F1 (95% CI)	Bal. Acc. (95% CI)
<i>Clinical-only (3 features: MMSE, CDR-SB, FAQ)</i>				
	Logistic Regression	0.9539 $\pm$ 0.0041	0.8510 $\pm$ 0.0114	0.8490 $\pm$ 0.0121
	Random Forest	0.9490 $\pm$ 0.0044	0.8574 $\pm$ 0.0119	0.8609 $\pm$ 0.0119
	SVM	0.9498 $\pm$ 0.0056	0.8430 $\pm$ 0.0105	0.8392 $\pm$ 0.0104
	XGBoost	0.9449 $\pm$ 0.0063	0.8589 $\pm$ 0.0125	0.8621 $\pm$ 0.0127
<i>Biomarker + Demographic + Genetic (8 features)</i>				
	Logistic Regression	0.7384 $\pm$ 0.0200	0.5560 $\pm$ 0.0263	0.5608 $\pm$ 0.0266
	Random Forest	0.7455 $\pm$ 0.0150	0.5608 $\pm$ 0.0145	0.5752 $\pm$ 0.0161
	SVM	0.7356 $\pm$ 0.0197	0.5548 $\pm$ 0.0238	0.5633 $\pm$ 0.0233
	XGBoost	0.7388 $\pm$ 0.0157	0.5654 $\pm$ 0.0177	0.5749 $\pm$ 0.0196
<i>Fusion (11 features: all variables combined)</i>				
	XGBoost	<b>0.9559 <math>\pm</math> 0.0046<sup>†</sup></b>	0.8519 $\pm$ 0.0101	0.8555 $\pm$ 0.0101
	Random Forest	0.9538 $\pm$ 0.0056	0.8467 $\pm$ 0.0125	0.8497 $\pm$ 0.0130
	Logistic Regression	0.9494 $\pm$ 0.0036	0.8399 $\pm$ 0.0088	0.8367 $\pm$ 0.0086
	SVM	0.9495 $\pm$ 0.0049	0.8345 $\pm$ 0.0124	0.8322 $\pm$ 0.0126

Results are mean  $\pm$  95% CI ( $= 1.96 \times \text{SD} / \sqrt{15}$ ) across 15 outer folds. Bold indicates numerically highest AUC. <sup>†</sup>Fusion XGBoost vs. Clinical-only Logistic Regression: Wilcoxon signed-rank test  $W = 42$ ,  $p = 0.33$ ; paired  $t$ -test  $t_{14} = 1.31$ ,  $p = 0.21$ . The difference is not statistically significant.

All four classifiers using the clinical-only feature set achieved AUC-OVR between 0.9449 and 0.9539. Logistic Regression performed best (0.9539  $\pm$  0.0041, Macro-F1 = 0.8510  $\pm$  0.0114). The biomarker-plus-demographic-genetic set achieved AUC-OVR of 0.7356–0.7455 and Macro-F1 of 0.55–0.57 across all classifiers. Adding clinical scales to form the fusion set restored AUC-OVR to the clinical-only range, and XGBoost reached the highest value (0.9559  $\pm$  0.0046). Neither the Wilcoxon signed-rank test ( $W = 42$ ,  $p = 0.33$ ) nor the paired  $t$ -test ( $t_{14} = 1.31$ ,  $p = 0.21$ ) detected a difference between the fusion and clinical-only models.

### 3.3. Pairwise Binary Classification

Table 3 presents pairwise binary classification results using the fusion feature set across all three diagnostic contrasts.

**Table 3.** Pairwise binary classification performance using the fusion feature set (11 features; repeated nested cross-validation, 15 outer folds). Note that the fusion set includes clinical assessment scales; results for the biomarker-plus-demographic-genetic set alone are provided in Supplementary Table S1.

Task	Classifier	AUC (95% CI)	F1 (95% CI)	Bal. Acc. (95% CI)
<i>CN vs. AD</i>				
	XGBoost	<b>1.0000 ± 0.0000</b>	0.9948 ± 0.0034	0.9957 ± 0.0029
	Random Forest	1.0000 ± 0.0001	0.9939 ± 0.0034	0.9951 ± 0.0028
	SVM	0.9997 ± 0.0003	0.9835 ± 0.0077	0.9863 ± 0.0063
	Logistic Regression	0.9998 ± 0.0003	0.9830 ± 0.0075	0.9835 ± 0.0072
<i>CN vs. MCI</i>				
	Logistic Regression	<b>0.9302 ± 0.0126</b>	0.8162 ± 0.0293	0.8527 ± 0.0229
	Random Forest	0.9291 ± 0.0106	0.8280 ± 0.0180	0.8680 ± 0.0154
	XGBoost	0.9282 ± 0.0119	0.8280 ± 0.0200	0.8680 ± 0.0169
	SVM	0.9271 ± 0.0154	0.8289 ± 0.0234	0.8667 ± 0.0193
<i>MCI vs. AD</i>				
	Random Forest	0.9719 ± 0.0075	0.9113 ± 0.0125	0.9045 ± 0.0134
	Logistic Regression	<b>0.9739 ± 0.0069</b>	0.9018 ± 0.0156	0.8986 ± 0.0158
	SVM	0.9722 ± 0.0063	0.8981 ± 0.0164	0.8921 ± 0.0177
	XGBoost	0.9682 ± 0.0070	0.9034 ± 0.0166	0.8972 ± 0.0163

Results are mean ± 95% CI across 15 outer folds. Bold indicates best AUC per task.

CN vs. AD discrimination achieved  $AUC \geq 0.9997$  across all classifiers. However, this near-perfect result is largely attributable to the extreme separation provided by CDR-SB and FAQ between CN and AD groups (see Section 4.3 for discussion of definitional circularity). MCI vs. AD AUC ranged from 0.9682 (XGBoost) to 0.9739 (Logistic Regression). CN vs. MCI was the most difficult task (Logistic Regression  $AUC = 0.9302 \pm 0.0126$ ). For comparison, using plasma biomarkers alone (without clinical scales), pairwise AUCs were 0.6972 (CN vs. MCI), 0.9153 (CN vs. AD), and 0.7588 (MCI vs. AD), indicating that the high pairwise performance in Table 3 is driven by the clinical scale components of the fusion set. Aggregated confusion matrices (Supplementary Table S5) confirm that MCI is the primary source of misclassification: the fusion Random Forest model achieved MCI recall of 73.0%, compared with 88.3% for CN and 91.4% for AD.

### 3.4. Plasma Biomarker Subset Analysis

Table 4 presents selected results from the systematic enumeration of all 15 plasma biomarker subsets, evaluated for the three-class (CN/MCI/AD) classification task using Random Forest with fixed background features.

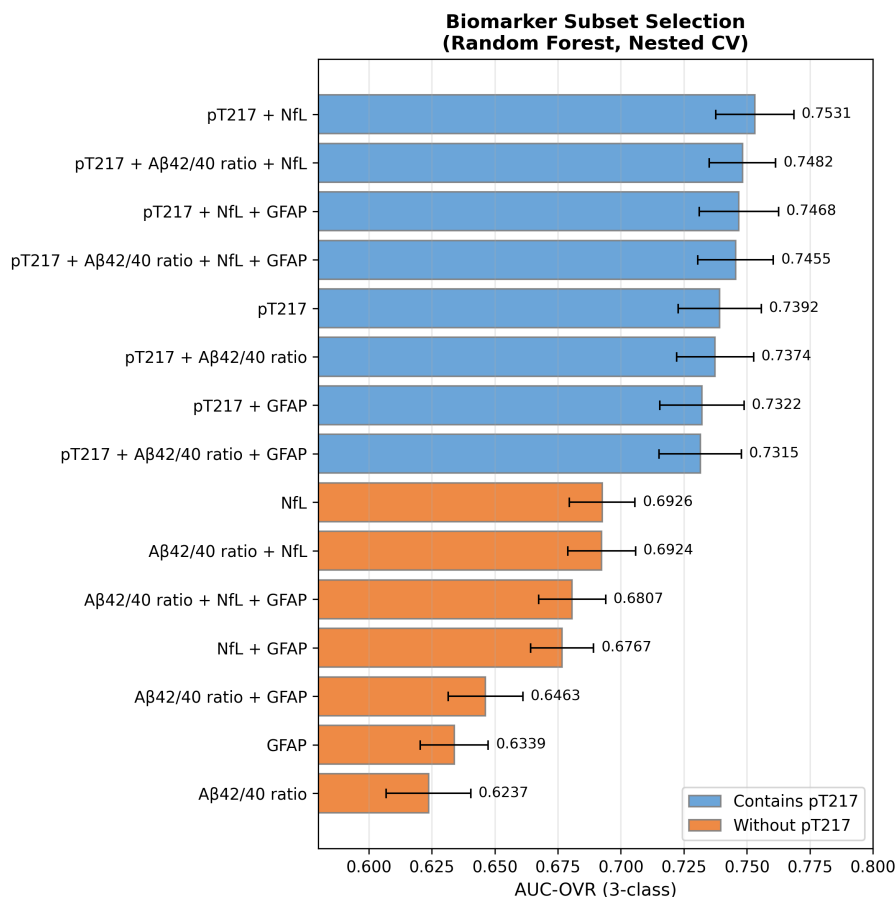
**Table 4.** Selected plasma biomarker subsets: three-class (CN/MCI/AD) AUC-OVR with fixed background features (Random Forest; 15 outer folds).

Biomarker Subset	CN/MCI/AD AUC-OVR
pT217 only	0.7392 ± 0.0165
A $\beta$ 42/40 only	0.6237 ± 0.0168
NfL only	0.6926 ± 0.0130
GFAP only	0.6339 ± 0.0135
<b>pT217 + NfL</b>	<b>0.7531 ± 0.0155</b>
pT217 + NfL + GFAP	0.7468 ± 0.0157
All four biomarkers	0.7455 ± 0.0150

Results are mean ± 95% CI (15 outer folds). Bold indicates the best-performing subset. Background features fixed in all models: age, sex, years of education, APOE4 allele count.

Three key findings emerged from the subset analysis. First, pT217 alone achieved the highest single-marker three-class AUC-OVR of  $0.7392 \pm 0.0165$ . Second, the two-marker combination pT217 +

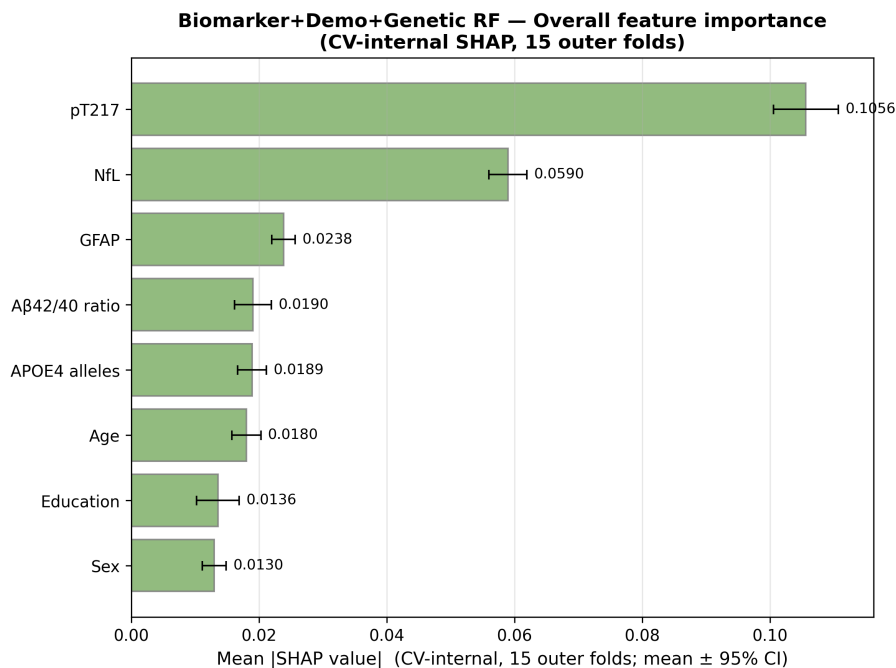
NfL was the best-performing subset (AUC-OVR =  $0.7531 \pm 0.0155$ ), followed by pT217 + NfL + GFAP (AUC-OVR =  $0.7468 \pm 0.0157$ ) and the full four-biomarker panel (AUC-OVR =  $0.7455 \pm 0.0150$ ). Third, adding A $\beta$ 42/40 to any subset did not improve and in some cases reduced AUC-OVR by 0.002–0.005, consistent with the small pairwise effect sizes of A $\beta$ 42/40 (one-way ANOVA  $p = 0.002$ ; Tukey HSD: AD vs. CN  $p = 0.002$ , but CN vs. MCI  $p = 0.18$  and MCI vs. AD  $p = 0.33$ ). Figure 2 displays the results for all 15 subsets: pT217-containing panels consistently outperformed those without pT217, while A $\beta$ 42/40-only subsets produced the lowest AUC-OVR values.



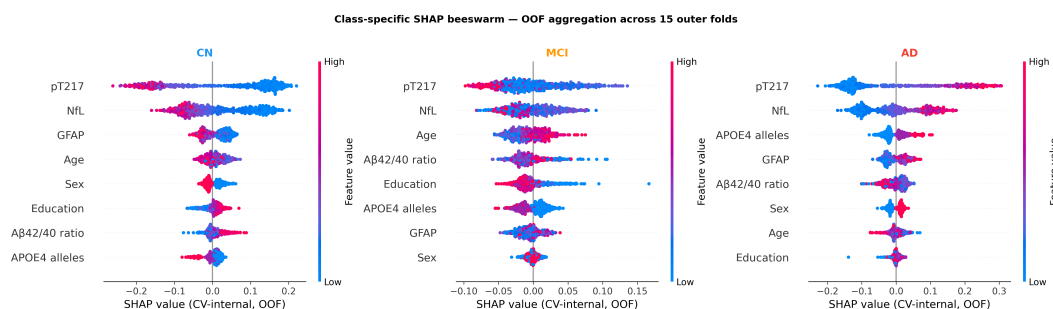
**Figure 2.** Biomarker subset selection results. Mean three-class AUC-OVR (repeated nested cross-validation, 15 outer folds) for all 15 non-empty subsets of {pT217, A $\beta$ 42/40, NfL, GFAP}, evaluated by Random Forest with fixed background features. The consistent dominance of pT217-containing panels and the limited contribution of A $\beta$ 42/40 are apparent.

### 3.5. Feature Interpretability via SHAP

To quantify each feature's contribution to individual predictions, SHAP values were aggregated across the 15 outer cross-validation folds (out-of-fold SHAP, mean  $\pm$  95% CI; see Methods §2.10). Each SHAP value quantifies how much a feature pushes a prediction toward or away from a particular class. Larger absolute values indicate stronger influence. Figure 3 shows that pT217 dominates global feature importance with a mean absolute SHAP of  $0.106 \pm 0.005$ , which is 1.8 times larger than NfL ( $0.059 \pm 0.003$ ). GFAP ( $0.024 \pm 0.002$ ), A $\beta$ 42/40 ( $0.019 \pm 0.003$ ), APOE4 ( $0.019 \pm 0.002$ ), age ( $0.018 \pm 0.002$ ), education ( $0.014 \pm 0.003$ ), and sex ( $0.013 \pm 0.002$ ) each contributed smaller effects. Figure 4 displays class-specific beeswarm plots: high pT217 values (red dots) shift leftward in the CN panel (decreasing CN probability) and rightward in the AD panel (increasing AD probability), with an attenuated pattern for MCI. Figure 5 decomposes mean absolute SHAP values by diagnostic class, showing that pT217 importance is highest for CN and AD but reduced for MCI, consistent with the difficulty of MCI detection.



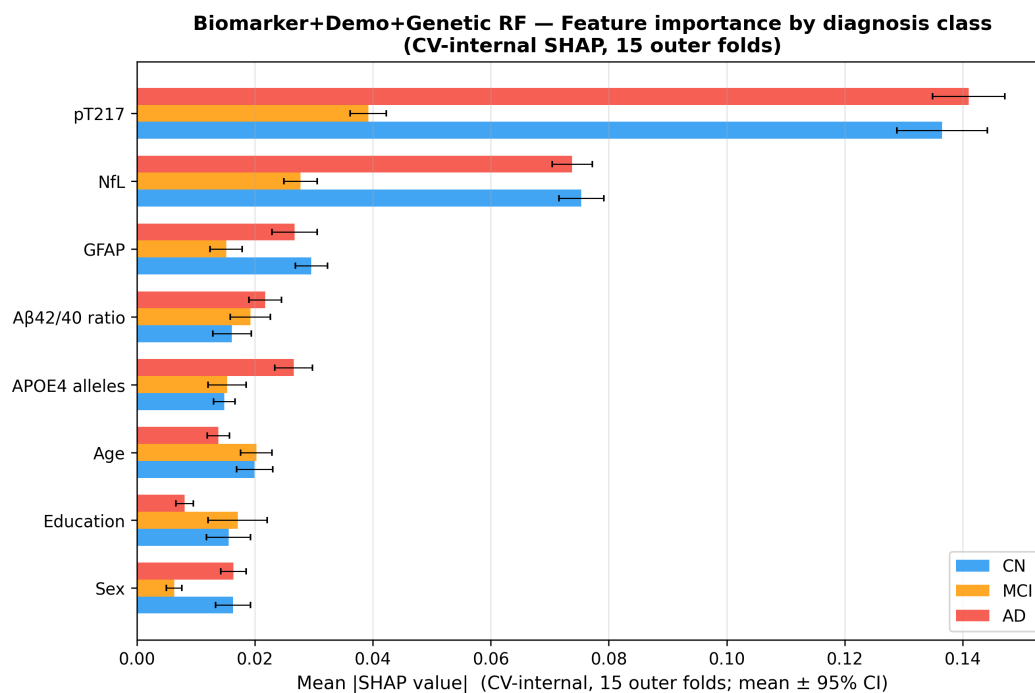
**Figure 3.** Overall feature importance from SHAP analysis (mean absolute SHAP value averaged across CN, MCI, and AD classes) for the biomarker-plus-demographic-genetic Random Forest model, aggregated across the 15 outer cross-validation folds (mean ± 95% CI; error bars). pT217 dominates global importance by a wide margin, followed by NfL, GFAP, Aβ42/40, and Age.



**Figure 4.** Class-specific SHAP beeswarm plots for CN (left), MCI (center), and AD (right). Each point represents a participant's out-of-fold SHAP value, averaged across the three outer test folds in which the participant appeared (15 outer folds total). pT217 is the dominant feature across all three classes: high pT217 values increase AD prediction probability and decrease CN probability.

Demographic variables (age, education, sex) and APOE4 allele count each contributed mean absolute SHAP values  $\leq 0.019$ , comparable to Aβ42/40 and notably smaller than pT217, NfL, and GFAP. Sex showed a larger SHAP contribution for CN than for MCI or AD. This class-specific effect requires replication in larger, sex-stratified cohorts.

To quantify the influence of definitional circularity on the fusion model, we applied the same SHAP analysis to the fusion Random Forest (11 features), aggregated across the 15 outer cross-validation folds. CDR-SB alone accounted for 41.7% of the total mean absolute SHAP value, followed by FAQ (23.0%) and MMSE (13.6%), for a combined 78.3% of the total feature importance from the three clinical scales. Among plasma biomarkers, pT217 was the largest contributor (9.1%), followed by NfL (4.6%), GFAP (2.2%), and Aβ42/40 (1.3%). Demographic and genetic variables (age, education, APOE4, sex) collectively accounted for 4.6%. These results demonstrate that the fusion model's performance is overwhelmingly driven by the clinical scales that participate in diagnostic label assignment, reinforcing the interpretation that the clinical-only AUC near 0.95 reflects definitional circularity rather than independent classification power (Supplementary Figure S1).



**Figure 5.** Mean absolute SHAP values by feature and diagnostic class (CN, MCI, AD), aggregated across the 15 outer cross-validation folds (mean  $\pm$  95% CI; error bars). pT217 and NfL show the highest class-differentiated importance.

### 3.6. External Validation in the CNTN Cohort

The reduced ADNI model transferred to CNTN with discrimination retained near the level achieved on ADNI itself with the same feature panel (Table 5). The best transferring model was the SVM, which achieved AUC-OvR = 0.702 (95% CI 0.635–0.764) on CNTN versus 0.687 in inner ADNI CV, indicating that discrimination was preserved when the ADNI-trained reduced model was applied to CNTN, i.e. no significant cross-cohort transfer loss under this matched feature panel. Per-class behaviour mirrored the patterns reported for ADNI in §3.3: CN was recovered with high recall (SVM: 77%, 44/57), while MCI again underperformed (27%, 15/56; Table 6), reproducing the same heterogeneity-driven pattern in an independent cohort.

**Table 5.** Zero-shot transfer of the ADNI-trained reduced biomarker model (NfL + GFAP + APOE + age + sex + education) to CNTN ( $n = 130$ : CN 57 / MCI 56 / AD 17). 95% CIs from 1000-iter bootstrap.

Model	ADNI inner CV AUC	CNTN AUC-OvR (95% CI)	Macro-F1 (95% CI)	Bal. Acc. (95% CI)	Brier (95% CI)
Logistic Regression	0.692	0.675 (0.603–0.741)	0.407 (0.334–0.484)	0.456 (0.360–0.552)	0.203 (0.188–0.217)
SVM (linear)	0.687	<b>0.702</b> <b>(0.635–0.764)</b>	0.460 (0.382–0.539)	0.503 (0.411–0.594)	0.197 (0.179–0.213)
Random Forest	0.680	0.645 (0.573–0.715)	0.400 (0.311–0.483)	0.470 (0.370–0.566)	0.204 (0.188–0.222)
XGBoost	0.666	0.623 (0.548–0.691)	0.409 (0.323–0.485)	0.464 (0.361–0.558)	0.213 (0.195–0.231)

**Table 6.** CNTN confusion matrix for the SVM (best transferring model).

True \ Predicted	CN	MCI	AD
CN	44	9	4
MCI	18	15	23
AD	2	7	8

## 4. Discussion

### 4.1. Comparison with Prior Studies

To situate our findings within the broader literature, Table 7 summarizes representative plasma biomarker and ADNI-based machine learning studies. Because these studies differ from ours in endpoint definition, input modality, and validation strategy, the comparison is intended to be contextual rather than a direct benchmark.

**Table 7.** Contextual comparison of selected plasma biomarker and ADNI classification studies, grouped by endpoint and modality.

Study	Cohort	Endpoint/Task	Modality	Validation	Performance
<i>Plasma biomarker studies (pathology endpoints)</i>					
Ashton et al. 2024 [7]	Multi-cohort ( $n = 786$ )	A $\beta$ -PET / tau-PET pathology positivity	Plasma pT217 (ALZpath)	Biomarker cutoff	0.92–0.97 (AUC)
Barthélemy et al. 2024 [18]	2 cohorts ( $n = 1,759$ )	A $\beta$ -PET / tau-PET pathology positivity	Plasma %p-tau217 (Mass spec)	Predefined cutoff	0.95–0.98 (AUC)
<i>Present study, ADNI (<math>n = 655</math>), repeated nested CV (15 outer folds)</i>					
Present study	ADNI ( $n = 655$ )	CN/MCI/AD	Plasma + demographics + APOE4	Repeated nested CV (RF)	0.7455 (AUC)
Present study	ADNI ( $n = 655$ )	CN/MCI/AD	Clinical scales	Repeated nested CV (LR)	0.9539 (AUC)
Present study	ADNI ( $n = 655$ )	CN vs. AD	pT217 + NFL + background	Repeated nested CV (RF)	0.9238 (AUC)
<i>ADNI machine learning classification studies</i>					
Cai et al. 2023 [21]	ADNI ( $n = 589$ )	4-yr conversion (CU / MCI strata)	Plasma p-tau + NFL + clinical + APOE4	Cross-validation	0.65 / 0.80 (AUC)
Bron et al. 2021 [19]	ADNI ( $n = 854$ )	CN vs. AD	Structural MRI	Nested CV (SVM)	0.940 (AUC)
Zhao et al. 2022 [20]	ADNI ( $n = 525$ )	CN vs. MCI / MCI vs. AD	Tau-PET radiomics	DLR + SVM	0.908/0.884 (Accuracy)

Studies are grouped by endpoint and modality. Background features = age, sex, education, APOE4 allele count. Abbreviations: LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; DLR, Deep Learning Radiomics; CV, cross-validation.

Three observations emerge from Table 7. First, the high AUCs of 0.92–0.98 reported by Ashton et al. [7] and Barthélemy et al. [18] were obtained against pathology-defined endpoints (amyloid-PET or CSF positivity) rather than clinical diagnostic categories. These studies establish that plasma p-tau217 tracks underlying AD pathology with high fidelity, and our SHAP results (in which pT217 dominates plasma biomarker importance) are consistent with this evidence. However, predicting pathology positivity is a fundamentally different and easier task than discriminating clinically heterogeneous CN, MCI, and AD groups, so the two sets of AUC values are not interchangeable measures of model utility.

Second, prior ADNI machine learning studies differ from the present work in both input modality and task formulation. Bron et al. [19] used structural MRI for binary CN vs. AD classification, Zhao et al. [20] used tau-PET radiomics for pairwise CN vs. MCI and MCI vs. AD tasks, and Cai et al. [21] addressed longitudinal conversion within fixed clinical strata. The present study uses peripheral plasma biomarkers together with demographics and APOE4 to classify all three clinical categories simultaneously, defining a distinct endpoint and modality combination relative to these prior reports.

Third, the evaluation protocol also differs. Most prior studies relied on a single train–test split or simple  $k$ -fold cross-validation, both of which are known to yield optimistic and high-variance estimates in small or class-imbalanced cohorts [12,13]. The repeated nested cross-validation protocol used here, 15 outer folds with three inner-fold repetitions, decouples hyperparameter tuning from performance estimation and yields more reliable estimates.

#### 4.2. SHAP Interpretability and Clinical Meaning

The SHAP analysis, aggregated across the 15 outer cross-validation folds (out-of-fold SHAP), confirms the biomarker subset ranking. pT217 achieved a mean absolute SHAP value of  $0.106 \pm 0.005$ , 1.8 times larger than NfL ( $0.059 \pm 0.003$ ). High pT217 values increased AD prediction probability and decreased CN probability, with an attenuated effect for MCI. This directional pattern is consistent with the known biology, as tau phosphorylation at threonine-217 correlates with amyloid plaque density and neurofibrillary tangle burden [7].

#### 4.3. Clinical Scales Dominate Three-Class Performance

A key interpretive constraint is the definitional circularity between clinical scale scores and ADNI diagnostic labels: MMSE, CDR-SB, and FAQ are themselves the principal instruments used to define diagnostic group membership [3]. Clinical diagnoses in ADNI incorporate CDR-global cutoffs directly (e.g., CDR-global = 0 for CN, CDR-global  $\geq 0.5$  for MCI, CDR-global  $\geq 1.0$  with cognitive impairment for AD), and MMSE and FAQ scores further inform the diagnostic algorithm. As a result, a substantial portion of the clinical-scale AUC of about 0.95 reflects this definitional overlap rather than independent predictive signal. This circularity means that clinical-only and biomarker-only AUC values reflect fundamentally different quantities and should not be directly compared. The biomarker AUC of 0.75 is free of this confound and therefore provides a more interpretable measure of genuine biological classification power.

Adding individual clinical scales to the biomarker set (Supplementary Table S2) showed that CDR-SB alone raised AUC from 0.7455 to 0.9502, FAQ alone to 0.8904, and MMSE alone to 0.8421. The fusion model (AUC = 0.9559, XGBoost) did not differ from clinical scales alone (AUC = 0.9539, Logistic Regression; Wilcoxon  $W = 42$ ,  $p = 0.33$ ; paired  $t$ -test  $t_{14} = 1.31$ ,  $p = 0.21$ ).

The fusion model SHAP analysis provides direct evidence of this circularity. CDR-SB alone contributed 41.7% of total feature importance, and the three clinical scales combined accounted for 78.3%, substantially exceeding all plasma biomarkers combined (17.1%) and demographic/genetic variables (4.6%). This confirms that the fusion model's performance is primarily attributable to the features used to define the diagnostic labels. Importantly, biomarker-only models provide an estimate of biological classification power that is free from this definitional confound, offering a more interpretable benchmark of genuine discriminative ability (AUC-OVR = 0.75).

#### 4.4. The Discriminative Value of Plasma Biomarkers

Within the modest three-class AUC of 0.74 achieved by the full plasma panel, subset analysis (Table 4) showed that pT217 alone already attained three-class AUC-OVR = 0.74, and the pT217 + NfL combination reached 0.75, indicating that pT217 carries virtually the entire discriminative signal of the plasma panel.

The CN vs. MCI task remained the most challenging for plasma biomarkers (pT217 + NfL three-class AUC = 0.75). The pattern of MCI misclassification is informative: aggregating the Random Forest biomarker-only confusion matrices across all 15 outer folds (168 MCI participants  $\times$  3 repeats = 504 MCI test instances), 244 (48.4%) were classified as CN, 160 (31.7%) as AD, and only 100 (19.8%) correctly identified as MCI. Errors fell predominantly toward the CN class, with 60.4% of errors being false negatives, consistent with the clinical overlap between early MCI and cognitively normal aging.

The A $\beta$ 42/40 ratio showed a significant but small between-group difference ( $p = 0.002$ ) and provided limited additional classification value beyond pT217-containing panels [5,6].

#### 4.5. APOE $\epsilon$ 4 and Genetic Risk Architecture

APOE4 carrier prevalence ranged from 36.1% (CN) to 64.4% (AD), consistent with prior genetic epidemiology of AD [14]. APOE4 homozygosity was present in 19.4% of AD participants versus 3.0% in CN.

#### 4.6. Methodological Strength and Rigour

This study applies repeated nested CV to a plasma biomarker classification problem and reports 95% confidence intervals across all metrics. The CNTN external evaluation (§3.6) provides additional evidence that the discriminative pattern is not specific to ADNI and partially generalises to an independent cohort with overlapping features.

#### 4.7. Limitations

Several limitations warrant consideration. The analytic sample ( $n = 655$ ) excluded 92 participants with missing biomarker data (ADNI sentinel values), who were on average younger, more likely female, and less likely to carry APOE4, suggesting potential selection bias toward a more clinically impaired sample. ADNI enrolls a predominantly non-Hispanic White, highly educated volunteer population from specialty memory clinics, which limits generalizability to community-based or ethnically diverse populations. Clinical-stage distributions in ADNI may not reflect the prevalence ratios encountered in primary care screening settings.

By design, this study focused exclusively on plasma-based and clinical measures, excluding neuroimaging data. While incorporating structural MRI or amyloid PET could potentially improve classification, the present pipeline prioritizes accessibility and cost-effectiveness for primary care screening settings. The CN vs. AD results ( $AUC \geq 0.9997$ ) using the fusion set are largely driven by clinical scales. However, the biomarker-plus-demographic-genetic set alone achieved CN vs. AD  $AUC = 0.9153 \pm 0.0134$  (Supplementary Table S1), confirming that CN and AD are also well-separated on biological grounds. Multiple pairwise and multiclass comparisons were conducted without formal correction (e.g., Bonferroni), so nominal  $p$ -values should be interpreted as exploratory. Multiclass and binary Brier scores are reported in Supplementary Tables S3 and S4. Calibration analysis (Supplementary Figure S2) showed good agreement across all three classes for the clinical-only and fusion models, whereas the biomarker-only model exhibited moderate miscalibration for MCI, consistent with the limited separability of this intermediate class using plasma markers alone. The present analysis is cross-sectional and therefore characterizes diagnostic group separation at a single time point rather than progression. The central findings of this work are properties of the feature space and are therefore independent of study design: the quantification of definitional circularity in clinical-scale-driven AUC, the identification of a plasma-alone three-class ceiling near 0.75, and the demonstration that pT217 carries virtually all of the plasma discriminative signal. Extending the framework to predict MCI-to-AD conversion in ADNI's longitudinal follow-up is a natural next step.

CNTN measures pTau-181 rather than pTau-217 and does not include plasma  $A\beta_{42/40}$ . Accordingly, the external validation in §3.6 was based on the feature intersection of the two cohorts (NfL, GFAP, APOE, age, sex, education), and strict replication of the full pT217-inclusive panel remains outstanding. Full external validation of the pT217-inclusive model is a priority for future work.

## 5. Conclusions

Clinical assessment scales (MMSE, CDR-SB, FAQ) achieved three-class AUC-OVR around 0.95 in the ADNI cohort, compared with 0.75 for plasma biomarkers augmented by demographic and genetic variables. The fusion of all feature types produced the highest three-class AUC (XGBoost:  $0.9559 \pm 0.0046$ ), matching the performance of clinical scales alone (Wilcoxon  $p = 0.33$ ). An important methodological consideration is that MMSE, CDR-SB, and FAQ overlap with the instruments used to assign ADNI diagnostic labels, so the clinical-only AUC partly reflects this definitional relationship. By contrast, plasma biomarker performance ( $AUC-OVR = 0.75$ ) is free of this confound and therefore provides a more meaningful measure of biological classification accuracy.

Among plasma biomarkers, pT217 achieved the highest single-marker three-class AUC-OVR (0.74), and pT217 + NfL was the best two-marker combination ( $AUC-OVR = 0.75$ ). External evaluation in the CNTN cohort ( $n = 130$ ) confirmed that the discriminative signal of the reduced biomarker panel transfers to an independent cohort ( $AUC-OVR = 0.702$  with the linear SVM), with the same

per-class pattern observed in ADNI: high CN recall and low MCI recall. MCI detection through plasma biomarkers alone remains limited.

Future directions should include: (1) validation in ethnically diverse, population-based cohorts; (2) longitudinal extension to ADNI follow-up data for MCI-to-AD conversion prediction; and (3) inclusion of newer generation biomarkers (e.g., pTau-231, synaptic proteins).

**Supplementary Materials:** The following supporting information can be downloaded at: [Preprints.org](https://preprints.org), Table S1: Pairwise binary classification AUC for all 15 non-empty plasma biomarker subsets; Table S2: Incremental contribution of clinical scales to three-class AUC-OVR; Table S3: Three-class multiclass Brier scores; Table S4: Pairwise binary Brier scores; Table S5: Aggregated confusion matrices for the best-performing model in each feature set; Figure S1: SHAP feature importance for the fusion Random Forest model; Figure S2: One-vs-rest calibration curves.

**Author Contributions:** Conceptualization, J.X. and F.C.; methodology, J.X.; software, J.X.; validation, J.X. and F.C.; formal analysis, J.X.; investigation, J.X.; resources, J.X.; data curation, J.X.; writing—original draft preparation, J.X.; writing—review and editing, J.X. and F.C.; visualization, J.X.; supervision, F.C.; project administration, F.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding. Funding sources for the underlying ADNI and CNTN data collection are detailed in the Acknowledgments.

**Institutional Review Board Statement:** The ADNI and CNTN studies were conducted in accordance with the Declaration of Helsinki and were approved by the Institutional Review Boards at all participating institutions, including the University of Southern California (USC) for the ADNI study protocols and the Cleveland Clinic Institutional Review Board for the CNTN study protocols. The present study is a secondary analysis of de-identified, publicly available data and did not require additional ethical approval.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the original ADNI and CNTN studies. Not applicable for the present secondary analysis.

**Data Availability Statement:** ADNI data are publicly available to qualified researchers at <https://adni.loni.usc.edu> upon completion of a data use agreement. CNTN data are available from the Center for Neurodegeneration and Translational Neuroscience (<https://nevadacntn.org/>). The Python analysis code is available from the corresponding author upon reasonable request.

**Acknowledgments:** Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<https://www.fnih.org>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
APOE4	Apolipoprotein E $\epsilon$ 4 allele
AUC	Area Under the Receiver Operating Characteristic Curve
AUC-OVR	AUC, One-vs.-Rest (multiclass)
A $\beta$ 42/40	Amyloid- $\beta$ 42/40 Ratio
CDR-SB	Clinical Dementia Rating Sum of Boxes
CN	Cognitively Normal
CNTN	Center for Neurodegeneration and Translational Neuroscience
CSF	Cerebrospinal Fluid
CV	Cross-Validation
FAQ	Functional Activities Questionnaire
GFAP	Glial Fibrillary Acidic Protein
LR	Logistic Regression
MCI	Mild Cognitive Impairment
ML	Machine Learning
MMSE	Mini-Mental State Examination
NfL	Neurofilament Light Chain
PET	Positron Emission Tomography
pT217	Plasma Phosphorylated Tau-217
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting

## References

1. World Health Organization. *Dementia: Key Facts*; WHO: Geneva, Switzerland, 2023. Available online: <https://www.who.int/news-room/fact-sheets/detail/dementia> (accessed on 1 March 2026).
2. Nichols, E.; Steinmetz, J.D.; Vollset, S.E.; et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050. *Lancet Public Health* **2022**, *7*, e105–e125. [https://doi.org/10.1016/S2468-2667\(21\)00249-8](https://doi.org/10.1016/S2468-2667(21)00249-8).
3. Jack, C.R.; Bennett, D.A.; Blennow, K.; et al. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* **2018**, *14*, 535–562. <https://doi.org/10.1016/j.jalz.2018.02.018>.
4. van Dyck, C.H.; Swanson, C.J.; Aisen, P.; et al. Lecanemab in Early Alzheimer's Disease. *N. Engl. J. Med.* **2023**, *388*, 9–21. <https://doi.org/10.1056/NEJMoa2212948>.
5. Hansson, O. Biomarkers for neurodegenerative diseases. *Nat. Med.* **2021**, *27*, 954–963. <https://doi.org/10.1038/s41591-021-01382-x>.
6. Barthelemy, N.R.; Horie, K.; Sato, C.; Bateman, R.J. Blood plasma phosphorylated-tau isoforms track CNS change in Alzheimer's disease. *J. Exp. Med.* **2020**, *217*, e20200861. <https://doi.org/10.1084/jem.20200861>.
7. Ashton, N.J.; Brum, W.S.; Di Molfetta, G.; et al. Diagnostic accuracy of a plasma phosphorylated tau 217 immunoassay for Alzheimer disease pathology. *JAMA Neurol.* **2024**, *81*, 255–263. <https://doi.org/10.1001/jamaneurol.2023.5319>.
8. Blennow, K.; Zetterberg, H. Biomarkers for Alzheimer's disease: Current status and prospects for the future. *J. Intern. Med.* **2018**, *284*, 643–663. <https://doi.org/10.1111/joim.12816>.
9. Benedet, A.L.; Mila-Aloma, M.; Vrillon, A.; et al. Differences Between Plasma and Cerebrospinal Fluid Glial Fibrillary Acidic Protein Levels Across the Alzheimer's Disease Continuum. *JAMA Neurol.* **2021**, *78*, 1471–1483. <https://doi.org/10.1001/jamaneurol.2021.3671>.
10. Ding, Y.; Sohn, J.H.; Kawczynski, M.G.; et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology* **2019**, *290*, 456–464. <https://doi.org/10.1148/radiol.2018180958>.

11. Battista, P.; Salvatore, C.; Berlingeri, M.; Cerasa, A.; Castiglioni, I. Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. *Neurosci. Biobehav. Rev.* **2020**, *114*, 211–228. <https://doi.org/10.1016/j.neubiorev.2020.04.026>.
12. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* **2018**, *180*, 68–77. <https://doi.org/10.1016/j.neuroimage.2017.06.061>.
13. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
14. Liu, C.C.; Kanekiyo, T.; Xu, H.; Bu, G. Apolipoprotein E and Alzheimer disease: Risk, mechanisms and therapy. *Nat. Rev. Neurol.* **2013**, *9*, 106–118. <https://doi.org/10.1038/nrneurol.2012.263>.
15. Alzheimer's Disease Neuroimaging Initiative (ADNI). Available online: <https://adni.loni.usc.edu> (accessed on 1 February 2026).
16. Center for Neurodegeneration and Translational Neuroscience (CNTN). 2015. Available online: <https://nevadacntn.org/> (accessed on 6 January 2025).
17. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
18. Barthélemy, N.R.; Salvadó, G.; Schindler, S.E.; et al. Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nat. Med.* **2024**, *30*, 1085–1095. <https://doi.org/10.1038/s41591-024-02869-z>.
19. Bron, E.E.; Klein, S.; Papma, J.M.; et al. Cross-cohort generalizability of deep and conventional machine learning for MRI-based diagnosis and prediction of Alzheimer's disease. *NeuroImage Clin.* **2021**, *31*, 102712. <https://doi.org/10.1016/j.nicl.2021.102712>.
20. Zhao, Y.; Zhang, J.; Chen, Y.; Jiang, J. A Novel Deep Learning Radiomics Model to Discriminate AD, MCI and NC: An Exploratory Study Based on Tau PET Scans from ADNI. *Brain Sci.* **2022**, *12*, 1067. <https://doi.org/10.3390/brainsci12081067>.
21. Cai, Y.; Fan, X.; Zhao, L.; et al. Comparing machine learning-derived MRI-based and blood-based neurodegeneration biomarkers in predicting syndromal conversion in early AD. *Alzheimers Dement.* **2023**, *19*, 4987–4998. <https://doi.org/10.1002/alz.13083>.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.