POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Multimodal Subject-Level Modeling for Alzheimer's Disease Classification

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Lorenzo Tanzi

Advisor: Prof. Lara Cavinato

Co-advisor: Prof. Chunfeng Lian (Xi'an Jiaotong University)

Academic year: 2025-2026

1. Introduction

Alzheimer's disease (AD) is the leading cause of dementia in older individuals and is clinically described as a progression along a continuum, conventionally discretised into three stages: cognitively normal (CN), mild cognitive impairment (MCI) and AD dementia [4]. The two endpoints are comparatively easy to separate, whereas the intermediate MCI class is biologically heterogeneous and overlaps, at the level of imaging and biomarker features, with both extremes. It is consequently the dominant source of error in three-class classification and, at the same time, the class of greatest clinical interest, since it is where early intervention is still plausible.

Imaging and biological biomarkers provide complementary windows on the pathological cascade. Within the NIA-AA ATN framework [2], the modalities used here probe distinct axes: Tau PET quantifies neurofibrillary tau deposition (T), structural MRI measures neurodegeneration as atrophy (N) and plasma biomarkers give minimally invasive systemic proxies of pathology, while region-of-interest (ROI) summaries provide a compact, interpretable representation of the normalised images. No single source is expected to resolve the three-class task on its own, which motivates multimodal fusion. Despite an extensive literature, the methodological foundations of this fusion are often weaker than the reported accuracies suggest. Three problems recur. The simultaneous CN/MCI/AD task is frequently reduced to easier binary comparisons, with the intermediate class set aside. Very high three-class accuracies are commonly accompanied by signs of optimistic evaluation and poor out-of-sample generalisation. The fusion itself is often performed without adequate control of information leakage, when the predictions feeding a meta-learner are produced on the same data used to train the base models, the combined estimate is inflated. A solid stacking procedure instead requires every fusion input to be an out-of-fold (OOF) prediction [5] and all branches to be aligned at the subject level under one identical set of cross-validation folds.

This thesis confronts these issues directly. It develops a leakage-controlled framework for subject-level CN/MCI/AD classification that fuses four branches: a 3D Tau PET model, a 3D structural MRI model, a tabular ROI-and-plasma model and a hierarchical PET-plasma model. All branches are derived from a cus-

tom, reproducible preprocessing and feature-extraction pipeline, aligned over 881 ADNI subjects [3] and evaluated under identical, no-leak folds. The contribution lies not in the backbone architectures, which follow established designs, but in the hierarchical formulation of the classification task and in a systematic, leakage-controlled comparison and fusion of the four modalities that quantifies the diagnostic information each one carries. The central empirical finding is that fusion improves significantly over every single branch, with the largest gain on MCI, yet MCI remains the limiting factor for reasons that lie in the information available rather than in the fusion mechanism.

2. Data and methodology

Cohort and leakage control. All data are drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [3]. Each subject is associated with a T1-weighted MRI, a Tau PET acquisition (tracer AV1451), plasma biomarkers and demographics. The final cohort comprises 881 subjects for which all branches are defined, with a marked class imbalance (512 CN, 272 MCI, 97 AD). Two design choices enforce a leakage-controlled evaluation and are the backbone of the whole work. First, every analysis is performed at the subject level. Scan-level probabilities are aggregated to a single vector per subject and partitions are defined over subjects, so that scans of the same subject never appear in both training and test. Second, all branches share a single set of synchronized five-fold partitions, defined once at the subject level by stratified group k-fold, so that the OOF predictions of the different branches refer to the same subjects under the same folds and can be safely combined.

Preprocessing pipeline. The preprocessing is MRI-centric. Structural volumes are conformed to 1 mm isotropic resolution, skull-stripped with SynthStrip [1], bias-corrected with FSL FAST and registered to MNI152 with an affine (12-DOF) transform. PET is rigidly registered to the subject's MRI and the PET-to-MRI and MRI-to-MNI transforms are composed into a single PET-to-MNI map applied in one resampling step,

$$T_{\mathrm{PET}\to\mathrm{MNI}} = T_{\mathrm{MRI}\to\mathrm{MNI}} \circ T_{\mathrm{PET}\to\mathrm{MRI}}, \qquad (1)$$

which avoids accumulating interpolation error. Harvard–Oxford ROI summaries are then extracted from the normalised volumes and combined with plasma markers and demographics into a subject-level tabular representation.

The four branches. Each branch b outputs a class-score vector $\mathbf{p}_s^{(b)}$ over $\{\mathrm{CN}, \mathrm{MCI}, \mathrm{AD}\}$. The PET-only and MRI-only branches are 3D DenseNet-121 classifiers implemented in MONAI, but trained with different objectives. The PET branch uses a class-balanced focal loss based on the effective number of samples, which concentrates the gradient on the hard, near-boundary subjects typical of the Tau signal. The MRI branch uses a milder class-weighted cross-entropy with a subject-balanced sampler, matched to its extreme-oriented signal. This asymmetry mirrors the asymmetry between the two modalities. The tabular branch ranks ROI features within each training fold by a univariate ANOVA F-statistic and feeds the top-k subset to an elastic-net/ExtraTrees ensemble, with all preprocessing fitted only on training subjects. The hierarchical branch performs an earlier, modality-level fusion of PET and plasma (one of the methodological contributions of this work) by decomposing the task into two binary stages, CN-vs-non-CN and MCI-vs-AD, each combining a PET model with a plasma logistic meta-model and recomposes the three-class vector as $P(\mathrm{CN}) = 1 - \hat{p}_1$, $P(\mathrm{MCI}) = \hat{p}_1(1 - \hat{p}_2)$, $P(\mathrm{AD}) = \hat{p}_1\hat{p}_2$.

Multimodal fusion. The four branches are combined using only their synchronized OOF predictions, so that the fusion never observes an in-sample prediction. The primary model is a fixed, TAB-anchored convex combination

$$\begin{aligned}\bar{\mathbf{p}}_s = {} & 0.55\,\mathbf{p}_s^{(\mathrm{TAB})} + 0.15\,\mathbf{p}_s^{(\mathrm{MRI})} \\ & + 0.15\,\mathbf{p}_s^{(\mathrm{PETONLY})} + 0.15\,\mathbf{p}_s^{(\mathrm{HIER})},\end{aligned} \qquad (2)$$

whose weights are fixed a priori from the relative strength of the branches and are not fitted on the evaluation data, which removes any risk of the fusion overfitting the folds and establishes a transparent, parameter-free decision rule. To test whether these weights leave performance on the table, a nested convex superlearner [5] instead learns the convex weights from the OOF

predictions, parametrised through a softmax and anchored towards a reference vector by a quadratic penalty of strength λ selected by nested cross-validation. A multinomial logistic meta-learner, trained on diagnostics derived from the branch outputs, provides a more expressive, MCI-oriented alternative. All models are evaluated on aggregated, OOF, subject-level predictions, with significance assessed by exact McNemar tests and 95% subject-level bootstrap confidence intervals.

3. Results

Across branches, the CN and AD classes are comparatively separable, whereas MCI is the hardest class for every model, its one-vs-rest AUC stays well below the AD one (0.84–0.92) and the dominant error mode is always the MCI→CN confusion. The branches are reliable in genuinely different regions and this non-overlap is what makes their fusion worthwhile (Table 1). The tabular branch is the strongest single model (accuracy 0.671, macro AUC 0.791, lowest fold-to-fold variability ±0.016). MRI is a CN-oriented detector (highest CN recall 0.857, lowest MCI recall 0.276). PET-only reaches the best AD recall (0.691) with a better MCI recall (0.397). The hierarchical branch attains the best single-branch MCI behaviour ($F1_{MCI} = 0.455$, MCI recall 0.474) and the best calibration (ECE = 0.040), at the price of the highest fold-to-fold variability, a direct consequence of its threshold-based two-stage rule.

Table 1: Single-branch and fusion performance on the synchronized 881-subject cohort (pooled OOF). Acc: accuracy; BAcc: balanced accuracy; mF1: macro-F1; mAUC: macro one-vs-rest AUC; $F1_M$/Rec_M: MCI F1 and recall.

Model	Acc	BAcc	mF1	mAUC	$F1_M$	Rec_M
PET-only	0.623	0.606	0.580	0.746	0.411	0.397
MRI-only	0.649	0.577	0.562	0.718	0.361	0.276
Tabular	0.671	0.605	0.590	0.791	0.400	0.338
Hierarchical	0.628	0.610	0.594	0.760	0.455	0.474
Fixed Fusion	**0.703**	**0.660**	**0.652**	0.804	**0.496**	0.452
MCI-orient.	0.672	0.614	0.630	**0.810**	0.507	**0.537**

The fusion improves on the best branch. The fixed convex fusion reaches accuracy 0.703, balanced accuracy 0.660 and macro-F1 0.652, against 0.671 and 0.590 for the strongest single branch. The gain is largest on the hardest class, with MCI F1 rising from 0.400 to 0.496 and MCI recall from 0.338 to 0.452. An exact McNemar test confirms a significant difference from every single branch (against tabular $p = 0.04$; against MRI $p = 0.009$; against PET-only and hierarchical $p < 10^{-7}$). The 95% bootstrap CIs for accuracy $[0.672, 0.732]$ and macro-F1 $[0.616, 0.688]$ exclude the single-branch values. The pooled confusion matrix (Table 2) shows the mechanism. The fusion draws CN stability from the MRI and tabular branches and gains MCI/AD sensitivity from the PET-based branches, shifting a substantial number of MCI subjects away from the CN prediction while keeping severe CN↔AD confusions low (12 subjects). The residual error remains the MCI→CN confusion (116 subjects).

Table 2: Pooled subject-level confusion matrix of the fixed convex fusion. Rows are true classes, columns predicted classes.

	$\widehat{\text{CN}}$	$\widehat{\text{MCI}}$	$\widehat{\text{AD}}$
CN	429	74	9
MCI	116	123	33
AD	3	27	67

The learned fusion reproduces the fixed one. A central methodological question is whether the a-priori weights $[0.55, 0.15, 0.15, 0.15]$ underperform weights learned from the data. The nested super-learner answers it empirically. When the convex weights are estimated from the OOF predictions, they converge back to the fixed configuration. In the two folds where a single global weight vector is selected, the learned weights are $[0.550, 0.151, 0.150, 0.149]$ and $[0.551, 0.148, 0.152, 0.149]$; in the three classwise folds the tabular branch stays dominant for every class ($w_{TAB} \in [0.53, 0.60]$). The resulting performance is statistically indistinguishable from the fixed fusion (accuracy 0.697 versus 0.703; $p = 0.18$ on only nine discordant subjects). The data-driven optimum therefore reproduces, rather than improves on, the a-priori weights, which both validates the fixed choice and justifies preferring it as the more explainable model.

An MCI-oriented operating point. Because misclassifying an MCI subject as cognitively normal is clinically costlier than the reverse, the logistic meta-learner with strong MCI class weights offers a complementary operating point. It raises MCI recall to 0.537 and MCI F1 to 0.507 (the best of all models), with the highest macro AUC (0.810) and the fewest severe CN↔AD confusions (5 versus 12), at the cost of about three accuracy points (0.672) and lower CN recall. The fixed fusion and the meta-learner are two points on the same accuracy–sensitivity frontier; the appropriate choice depends on the clinical cost assigned to missed MCI cases.

4. Discussion

Why the branches complement each other. The fusion gain is not the fusion tracking its strongest input. Each branch is reliable in a different region of the problem and the asymmetric training choices are what make this complementarity usable. The mild MRI reweighting reflects a structural signal informative mainly for the extremes of the spectrum, where a stronger minority-focused objective destabilised an already CN-oriented branch. The harder PET focal loss reflects a Tau signal whose discriminative information is concentrated in a small number of ambiguous subjects. The hierarchical branch, designed around the asymmetric structure of the task, contributes an MCI-oriented error profile that differs from the flat classifiers, which is what makes it valuable in the combination rather than as a standalone model.

The fusion is close to its ceiling. Multimodal performance sits near the limit extractable from these branches by fusion alone. Learning the weights did not help, calibrating the meta-learner did not help (both Platt and isotonic calibration roughly halved MCI recall, since calibration fitted on the small inner folds is dominated by the majority class) and enriching the meta-learner with disagreement features did not improve on the simple convex combination. No configuration simultaneously achieves high accuracy and high MCI recall. The residual MCI difficulty is therefore not a deficiency of the fusion mechanism but a limit of the information content of the branches themselves, so further gains must come from new information, not from a more elaborate combination of the existing signals.

MCI as the discretisation of a continuum. Every branch and the fusion fail predominantly in the same direction, towards CN, because early MCI and normal ageing form a near-continuum and a substantial fraction of MCI subjects carry little detectable pathological signal at the time of acquisition. The CN/MCI/AD labels are partly arbitrary discretisations of a continuous process. Forcing a hard assignment penalises a model equally for a near-boundary CN/MCI confusion (clinically minor) and for a CN/AD confusion (clinically severe). Much of the error of the models comes from cutting a continuum into three boxes exactly where the biology is most continuous.

Limitations. The results rest on a single, class-imbalanced cohort. The small AD group drives the high fold-to-fold variability of the molecular and hierarchical branches and external validation on an independent dataset (e.g. OASIS) remains necessary before any claim of generalisation. The leakage-controlled, subject-level, synchronized-fold protocol strengthens internal validity but cannot substitute for that external check.

5. Future directions

The finding that MCI is an information limit rather than a modelling artefact shows how the most promising improvement is a reformulation of the task that respects the continuous nature of disease progression. The most consequential step recasts classification as regression on a continuous severity axis, predicting a graded disease-severity score (a cognitive or biomarker-derived progression index, or a data-driven latent axis) together with its uncertainty. Such a target dissolves the artificial sharpness of the CN/MCI boundary that drives most of the residual error and yields a clinically richer output, while the discrete decision remains recoverable by thresholding under costs that reflect the asymmetry between error types. A less radical intermediate is ordinal classification, which keeps the familiar labels but penalises errors by their distance along CN $<$ MCI $<$ AD, so that a CN/AD confusion becomes strictly more costly

than a CN/MCI one and the objective aligns with a clinical cost structure that the present metrics can only report after the fact. A closely related variant treats MCI as a transition region along a latent CN-to-AD continuum rather than as a third homogeneous class, interpreting it as the high-uncertainty band that the empirical behaviour observed here already suggests.
Two further directions address the information limit directly. Longitudinal information, such as the rate of hippocampal atrophy or of Tau accumulation between visits, carries far more about transition than a single snapshot and fits naturally into a regression-on-severity formulation. In the same spirit, biomarkers orthogonal to those used here, in particular cerebrospinal-fluid measures, could supply the signal the cross-sectional imaging modalities lack for the intermediate class. Finally, stability-oriented training, for example multi-seed ensembling of the base branches before fusion, offers a low-risk way to reduce the fold-to-fold variance of the molecular and hierarchical branches.

6. Conclusions

This work established a rigorous, leakage-controlled framework for subject-level CN/MCI/AD classification, fusing four complementary branches over a synchronized cohort of 881 ADNI subjects with all fusion inputs constrained to be synchronized out-of-fold predictions. The multimodal fusion improved measurably and significantly over every single branch, with the largest gain on the most difficult class. The fixed, parameter-free convex combination served as the primary model and a nested superlearner confirmed its weights rather than improving on them.
The central and persistent finding is that MCI remains the bottleneck because of the cross-sectional information available rather than any deficiency in the fusion mechanism. The way forward therefore lies in a different formulation of the task, along the continuous-progression directions outlined above, to which the leakage-controlled, subject-level protocol established here can be directly extended.

References

[1] Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: Skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022.

[2] Clifford R. Jack, David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. Nia-aa research framework: Toward a biological definition of alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[3] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner. Alzheimer's disease neuroimaging initiative (adni): Clinical characterization. *Neurology*, 74(3):201–209, 2010.

[4] Amir Abbas Tahami Monfared, Michael J. Byrnes, Leigh Ann White, and Quanwu Zhang. Alzheimer's disease: Epidemiology and clinical progression. *Neurology and Therapy*, 11:553–569, 2022.

[5] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6:Article 25, 2007.