uc3m | Universidad Carlos III de Madrid

Master's Degree in Machine Learning for Health

2025-2026

Master Thesis

"Long-Term Prediction of Alzheimers Disease Progression from MCI Using Combined 18F-FDG and Tau PET Imaging with Explainable Deep Learning"

Duarte Pinto Correia de Moura

David Izquierdo García

Leganés, February 2026

Long-Term Prediction of Alzheimer's Disease Progression from MCI Using Combined 18F-FDG and Tau PET Imaging with Explainable Deep Learning

Duarte Moura
Master in Machine Learning for Health
Universidad Carlos III de Madrid
Madrid, Spain

David Izquierdo García
Distinguished Professor
Universidad Carlos III de Madrid
Madrid, Spain

***Abstract*—Predicting the progression from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) is a major clinical challenge. While ^{18}F-FDG-PET is a widely available proxy for neurodegeneration, tau PET tracers offer pathology-specific information yet are severely limited by data scarcity. We propose a hierarchical multi-stage 3D CNN framework that leverages cross-modality transfer learning to bridge this gap. By transferring neurodegenerative representations learned from a large metabolic dataset (FDG) to the pathological protein domain (^{18}F-flortaucipir), our Tau model achieves comparable classification accuracy to the FDG model despite a threefold reduction in training data. To support clinical trust, we integrate Monte Carlo Dropout for uncertainty estimation and Grad-CAM for interpretability. Our results demonstrate that this approach enables reliability-aware risk stratification, where high-confidence predictions achieve approximately 90% accuracy for FDG-PET and 82% for Tau-PET.**

***Index Terms*—Alzheimer's Disease, MCI conversion, 3D CNN, transfer learning, FDG-PET, Tau-PET, Monte Carlo Dropout, explainable AI**

I. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder characterized by a long preclinical phase during which pathological changes accumulate before the onset of dementia. Mild Cognitive Impairment (MCI) represents a critical intermediate stage, where patients exhibit measurable cognitive decline while largely preserving functional independence. Identifying individuals with progressive MCI (pMCI)—those who will convert to AD—from stable MCI (sMCI) remains a major clinical challenge, with important implications for patient management, prognosis, and clinical trial stratification.

Biomarker-based approaches play a central role in addressing this challenge. Cerebrospinal fluid (CSF) markers and Positron Emission Tomography (PET) imaging provide complementary views of the disease process. Among PET modalities, ^{18}F-Fluorodeoxyglucose (FDG) PET is widely used as a marker of cellular metabolism that can be linked to synaptic dysfunction and neurodegeneration, but its patterns are not entirely specific to AD. In contrast, ^{18}F-Tau-PET tracers (e.g., AV-1451) enable direct visualization of neurofibrillary tau tangles, which follow a stereotypical spatiotemporal progression described by Braak staging and show strong associations with cognitive decline [3]. Despite its biological specificity, Tau-PET remains far less available than FDG-PET, resulting in limited sample sizes that pose a significant challenge for data-hungry deep learning models.

Recent advances in deep learning have led to promising results in AD-related imaging tasks, shifting from cross-sectional diagnosis toward the more demanding problem of predicting MCI-to-AD conversion. Santangelo et al. [4] showed that FDG-PET can reliably identify "imminent" converters within short time horizons (1–2 years), although long-term prediction remains difficult. More complex approaches, such as the multimodal radiomics framework proposed by Zhou et al. [5], have achieved high performance by combining longitudinal imaging, handcrafted features, and clinical data. While effective, these methods often require extensive multimodal inputs or longitudinal follow-up, limiting their applicability in routine clinical settings.

Alongside predictive performance, model interpretability and reliability have emerged as critical requirements for clinical adoption. Explainable deep learning approaches have been proposed to highlight disease-relevant regions in PET images, particularly for FDG-PET [6]. However, many existing models provide only point estimates, offering limited insight into prediction confidence and making it difficult to distinguish reliable predictions from uncertain ones. Moreover, most prior work focuses on a single imaging modality, without explicitly addressing the challenge of transferring knowledge between modalities with very different biological meanings and data availability.

In this context, there is a clear need for models that (i) leverage large, well-established FDG-PET datasets to mitigate data scarcity in Tau-PET, (ii) remain deployable using a single imaging modality at inference time, and (iii) provide both interpretability and uncertainty estimates to support clinical decision making.

A. Contributions

This study extends the work of Fernandez-Garcia et al. [1] by introducing the following contributions:

1) Cross-Modality Transfer Learning: : We demonstrate that a 3D CNN trained on FDG-PET can be successfully adapted to Tau-PET for MCI conversion prediction, achieving robust performance despite substantially reduced data availability.

2) Uncertainty Quantification: : We incorporate Monte Carlo Dropout to derive a Confidence Score, enabling reliability-aware stratification of predictions rather than relying solely on point estimates.

3) Biological Validation and Explainability: : We validate model predictions against established clinical (MMSE/MoCA) and genetic (APOE4) markers, and employ Grad-CAM to visualize anatomically meaningful decision regions consistent with known AD pathology.

II. Materials and Methods

To address the challenge of cross-modality prediction with limited data, our methodological framework proceeds in three stages: dataset curation, hierarchical model training with domain adaptation, and post-hoc reliability analysis.

A. Dataset Description

All data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The dataset composition is illustrated in Fig. 1.

For Stage 1 foundation model training, we used FDG-PET scans from subjects with definitive AD or CN diagnoses. The AD/CN cohort comprised 681 unique subjects: 306 AD subjects (539 scans) and 375 CN subjects (950 scans). Multiple longitudinal scans per subject were included when available.

For Stage 2 MCI conversion prediction, subjects were labeled based on longitudinal clinical follow-up. Subjects were classified as progressive MCI (pMCI) if they were diagnosed as MCI at baseline and converted to AD at any subsequent follow-up visit within 24 months. Subjects were classified as stable MCI (sMCI) if they remained diagnosed as MCI for at least 24 months without conversion. Subjects who reverted to CN, lacked sufficient follow-up, or exhibited ambiguous diagnostic trajectories were excluded.

The FDG-MCI cohort consisted of 627 subjects: 298 pMCI and 329 sMCI. Multiple longitudinal scans per subject were included when available. The Tau-PET cohort, constrained by tracer availability, comprised 200 subjects: 75 pMCI and 125 sMCI. For Tau-PET, one baseline scan per subject was used to maintain consistency and avoid temporal confounds.

For FDG-PET, all available scans per subject were included to maximize data usage, while enforcing strict subject-level separation across cross-validation folds such that no subject contributed scans to more than one fold. For Tau-PET, only the earliest available baseline scan per subject was used, reflecting the limited longitudinal availability of tau imaging and ensuring consistent temporal alignment with baseline clinical labels.

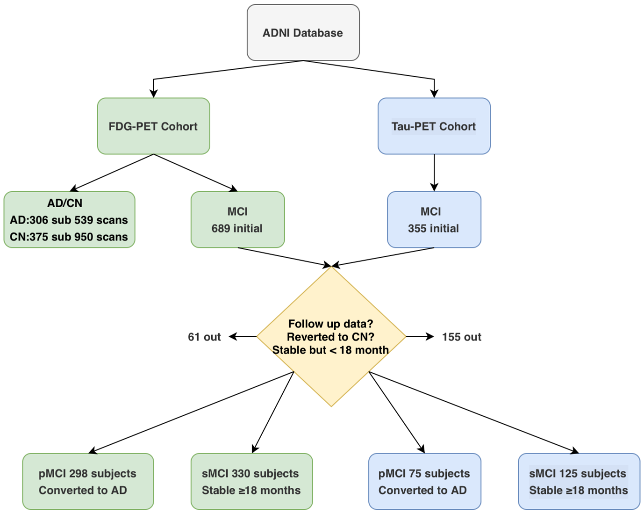


Fig. 1: Dataset composition for AD/CN classification and MCI conversion prediction across FDG-PET and Tau-PET modalities.

B. Preprocessing Steps

Preprocessing was performed in two stages: external standardization by ADNI and internal processing for deep learning readiness.

1) ADNI Standardization: Images retrieved from ADNI underwent the standard pre-processing protocol [7]. Each tracer type's first scan was reoriented into a standard $160 \times 160 \times 96$ voxel image grid with $1.5\,\text{mm}^3$ isotropic voxels using rigid body registration to achieve standard AC-PC alignment. Subsequent scans were co-registered to the baseline AC-PC alignment.

2) Internal 3D Pipeline and Skull-Stripping Strategy: All volumes were resampled to $100 \times 100 \times 90$ using trilinear interpolation and normalized using per-volume min–max scaling.

Although tau tracers such as ^{18}F-flortaucipir (AV-1451) are known to exhibit extracerebral uptake in skull/bone and meningeal structures [9]–[12], we evaluated two preprocessing approaches: (1) retaining the full field-of-view and relying on the CNN to suppress extracerebral signal, and (2) applying a robust deep learning-based skull-stripping method prior to model training. Preliminary experiments with conventional skull-stripping methods (e.g., Otsu thresholding combined with 3D connected component analysis) confirmed reports in the literature that off-target binding can reduce masking robustness, leading to partial cortical signal loss or inconsistent boundaries across subjects.

To address this challenge, we applied SynthStrip [8], a deep learning skull-stripping tool trained on synthesis-based domain randomization that generalizes robustly across imaging modalities, resolutions, and populations (Fig. 2). SynthStrip successfully extracted clean brain masks from both FDG-PET and Tau-PET scans without cortical signal loss, enabling uniform preprocessing across modalities. Comparative experiments demonstrated that skull-stripped images yielded substantially improved model performance and anatomical in-

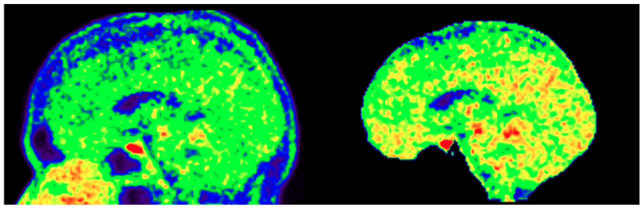


Fig. 2: Comparison of Tau-PET images before (left) and after (right) SynthStrip skull-stripping, demonstrating clean removal of extracerebral uptake while preserving cortical signal integrity.

terpretability compared to full field-of-view inputs, particularly for Tau-PET. Consequently, all FDG-PET and Tau-PET scans underwent identical preprocessing, including SynthStrip-based skull-stripping, prior to model training and inference.

C. Implementation Details

We employed a shallow 3D Convolutional Neural Network (CNN) designed to mitigate overfitting, a common challenge in medical imaging where the number of features far exceeds the number of training subjects. The network consists of four convolutional blocks, each composed of $3 \times 3 \times 3$ convolutions followed by ReLU activation, Batch Normalization, and Max Pooling. Dropout is applied after each block to improve regularization. The feature extractor is followed by a fully connected classification head with 768 units.

Model training followed a hierarchical multi-stage procedure, illustrated in Fig. 3. In Stage 1, the network was trained to discriminate AD from CN subjects using FDG-PET, allowing the convolutional layers to learn general patterns of neurodegeneration and brain anatomy. In Stage 2, the model was adapted to the MCI conversion task using FDG-PET, with the classifier head re-initialized to focus on the subtler differences between progressive and stable MCI. In Stage 3, the learned representations were transferred to Tau-PET for MCI conversion prediction. To reduce the impact of cross-modality domain shift and catastrophic forgetting, Batch Normalization statistics were frozen and a low learning rate ($OriginalLR/20$) was used during this final adaptation stage.

All MCI conversion experiments were conducted using 5-fold cross-validation with strict subject-level splitting to prevent data leakage. Although multiple FDG-PET scans were available for some subjects, no subject appeared in more than one fold. For Tau-PET, one baseline scan per subject was used. Model selection within each fold was based on validation loss. While subject overlap exists between the pretraining (AD/CN) and fine-tuning (MCI) cohorts, data leakage is mitigated by the distinct tasks (diagnosis vs. prognosis) and the temporal separation of scans. Furthermore, strict subject-level separation was maintained within the cross-validation folds of the MCI experiments.

To further mitigate overfitting and improve generalization, on-the-fly data augmentation was applied during training. Augmentations were implemented using the MONAI framework and executed directly on the GPU to minimize computational overhead. Each training volume was randomly transformed with a probability of 0.5 using a combination of spatial and intensity perturbations, including 3D affine transformations (random rotations up to approximately 20° and translations). Data augmentation was applied consistently across FDG-PET and Tau-PET experiments and did not alter the underlying anatomical structure of the images, serving instead to promote robustness to small spatial misalignments and scanner-related variability.

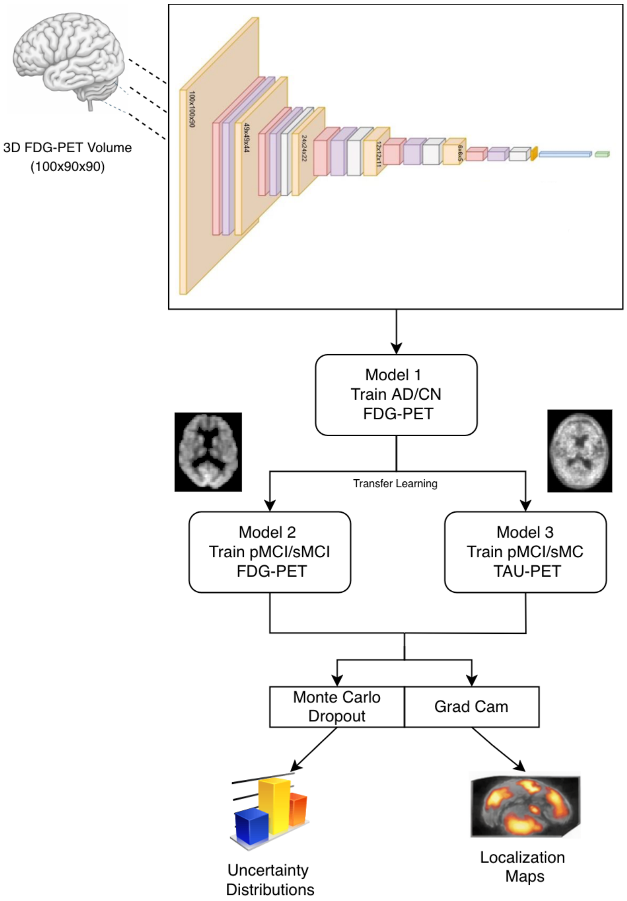


Fig. 3: Hierarchical multi-stage training pipeline with transfer learning from AD/CN foundation model to FDG and Tau MCI conversion predictors.

The final architecture and training hyperparameters were selected through a systematic search using the Optuna framework. Optimization trials were conducted, exploring the number of convolutional layers, base filter width, dropout rate, and size of the classification head. The selected configuration represents a balance between model capacity and generalization, and was fixed for all subsequent experiments.

D. Uncertainty Estimation and Explainability

1) Uncertainty estimation: To quantify predictive uncertainty, Monte Carlo Dropout was applied at inference time [16] by keeping dropout layers active and performing 500 stochastic forward passes per scan. This procedure yields a distribution of predicted probabilities for each subject. The standard deviation (SD) of this distribution was used as a measure of uncertainty. A Confidence Score (CS) was then defined as:

$$CS = 100 - SD_{\text{percentile}} \tag{1}$$

where the percentile is computed relative to the SD distribution of the validation set within each cross-validation fold and imaging modality. This normalization allows confidence values to be compared across folds and modalities.

The standard deviation percentile was computed independently within each cross-validation fold and imaging modality to account for differences in predictive variability across training splits and domains. Percentile-based normalization was adopted to yield a comparable Confidence Score scale across FDG-PET and Tau-PET despite differences in absolute uncertainty magnitude.

2) Explainability: Model interpretability was assessed using Gradient-weighted Class Activation Mapping (Grad-CAM) [2]. Grad-CAM generates coarse localization maps by backpropagating gradients from the output logit corresponding to the pMCI class to the final convolutional layer, highlighting spatial regions that most strongly influence the model's decision. The resulting volumetric heatmaps were upsampled to the input resolution for visualization and subsequent analysis.

In this work, we adopt post-hoc interpretability methods rather than intrinsically self-explainable network architectures. While self-explainable models offer theoretical advantages by embedding interpretability directly into the model structure, they typically impose architectural constraints that can reduce expressive capacity and predictive performance, particularly in high-dimensional imaging settings. Prior studies have highlighted this trade-off, noting that enforcing interpretability at the architectural level often limits model flexibility and optimization [13], [14]. In clinical neuroimaging applications, spatial attribution—namely understanding *where* in the brain a model bases its prediction—is often more actionable than interpreting internal feature representations. Given the performance-critical nature of MCI-to-AD prognosis, we therefore prioritize predictive accuracy while providing clinically meaningful spatial explanations through post-hoc attribution methods such as Grad-CAM.

III. Results

A. Cohort Characteristics and Baseline Performance

Table I summarizes baseline demographic and clinical characteristics of the FDG-MCI and Tau-MCI cohorts at the subject level. The two cohorts are well matched in terms of age, education, sex distribution, and baseline cognitive scores (MMSE), reducing the likelihood that observed performance differences are driven by cohort imbalance rather than imaging modality.

The Stage 1 FDG-based foundation model achieved a validation loss of 0.2870 on the AD vs. CN task (Val Acc: 0.8986, Val AUC: 0.9571), indicating stable convergence and providing a strong initialization for subsequent MCI conversion training.

TABLE I: Cohort Demographics and Composition

Variable	FDG-PET	Tau-PET
Stage 1: AD/CN Classification		
Total Subjects	681	—
AD Subjects	306	—
CN Subjects	375	—
Stage 2: MCI Conversion Prediction		
Total MCI Subjects	627	200
pMCI	298	75
sMCI	329	125
Baseline Demographics (MCI Cohort)		
Age (Mean ± SD)	73.1 ± 7.3	72.0 ± 7.4
Education (Years)	16.0 ± 2.7	16.1 ± 2.7
MMSE Score	27.6 ± 1.7	27.8 ± 1.8
Female (%)	40.9%	40.5%

TABLE II: MCI-to-AD Classification Performance (5-Fold Cross-Validation)

Modality	Mean AUC-ROC	Mean Accuracy
MCI FDG-PET	0.791 ± 0.041	0.697 ± 0.043
MCI Tau-PET	0.773 ± 0.067	0.705 ± 0.048

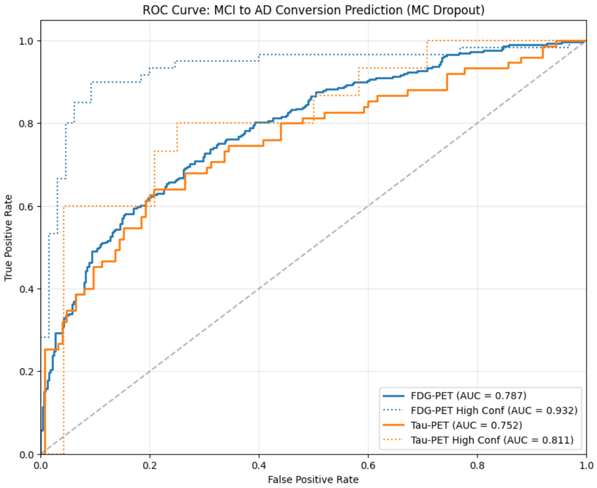


Fig. 4: ROC curves for MCI-to-AD conversion prediction computed at inference time using Monte Carlo Dropout. For each subject, the predicted probability is taken as the mean of the stochastic forward passes.

B. Prediction Performance and Uncertainty

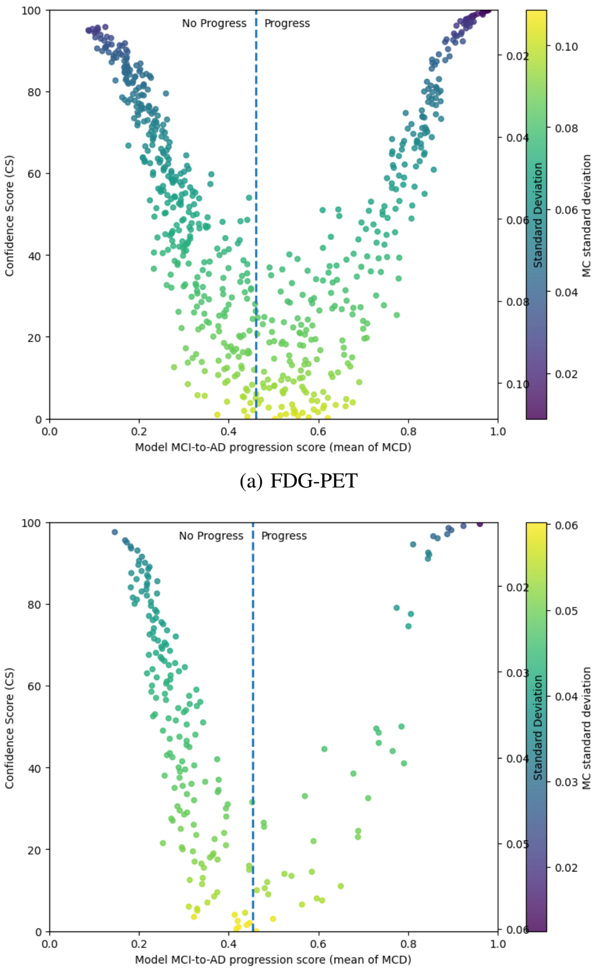
As detailed in Table II, both modalities achieved comparable mean accuracy across 5-fold cross-validation, while FDG-PET yielded a higher mean AUC-ROC than Tau-PET. Fig. 4 reports the corresponding ROC curves computed using the mean Monte Carlo Dropout prediction per subject.

To examine performance as a function of prediction confidence, Table III reports AUC and accuracy after stratifying subjects into three Confidence Score (CS) bins. For both modalities, performance differs across CS strata, with the highest accuracy observed in the high-confidence bin and the lowest accuracy observed in the low-confidence bin.

Figs. 5a and 5b visualize the relationship between the mean MCI-to-AD progression score and the Confidence Score at

TABLE III: Performance Stratification by Confidence Score

CS Bin	FDG-PET		Tau-PET	
	AUC	Acc	AUC	Acc
Low (<40)	0.692	64.4%	0.681	65.0%
Mid (40–80)	0.729	68.9%	0.799	76.2%
High (>80)	**0.932**	**89.6%**	**0.811**	**82.1%**



(a) FDG-PET

(b) Tau-PET

Fig. 5: Relationship between mean MCI-to-AD progression score (mean of Monte Carlo Dropout predictions) and Confidence Score (CS) for FDG-PET and Tau-PET. The vertical dashed line denotes the decision threshold; point color encodes the raw Monte Carlo standard deviation, while CS corresponds to the percentile-normalized inverse of this variability.

the subject level for FDG-PET and Tau-PET, respectively. In both modalities, confidence values are higher at more extreme predicted scores and tend to decrease closer to the decision threshold. Additionally, subjects with similar mean scores can exhibit different confidence levels, particularly near the threshold region. This subject-level behavior is consistent with the performance stratification by Confidence Score shown in Table III.

TABLE IV: Spearman Correlations (ρ) with Clinical Variables

Modality	MMSE	MoCA	APOE4	Age
FDG-PET	-0.259***	-0.325***	0.211***	0.156***
Tau-PET	-0.183*	-0.212*	0.217*	0.146*

Note: * $p < 0.05$, *** $p < 0.001$

C. Clinical and Biological Validation

To assess whether model predictions aligned with established disease markers, predicted conversion risk scores were correlated with clinical and genetic variables (Table IV). For both modalities, higher predicted risk was associated with lower cognitive performance, as evidenced by significant negative correlations with MMSE and MoCA scores. In addition, positive correlations with APOE4 carrier status were observed, indicating that the model captures known genetic risk factors [17].

Correlations were generally stronger for FDG-PET than for Tau-PET, particularly with cognitive measures.

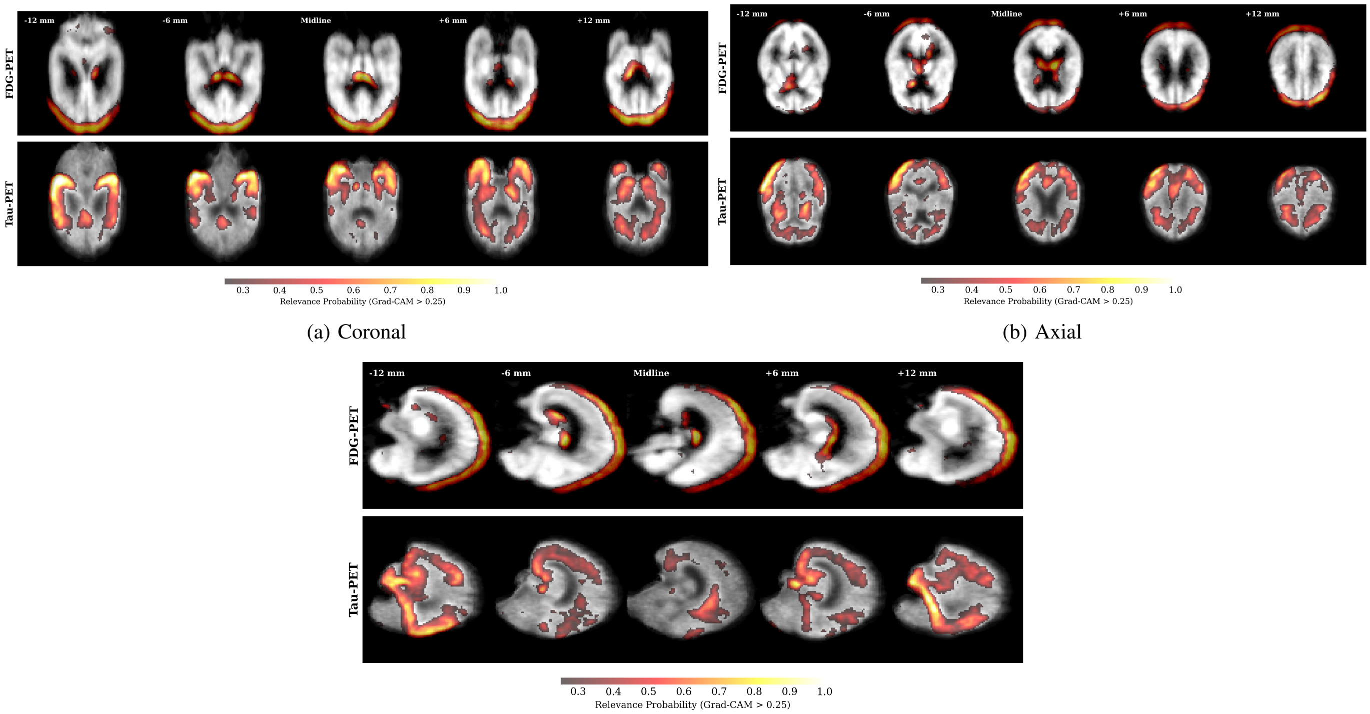
D. Explainability and Neuroanatomical Patterns

Grad-CAM saliency maps were generated for a subset of high-confidence pMCI predictions (model probability > 0.80) to examine the neuroanatomical regions contributing most strongly to model decisions. Grad-CAM was computed with respect to the pMCI output logit at the final convolutional block, and the resulting attribution maps were upsampled to the input resolution for visualization. Representative examples across anatomical planes are shown in Fig. 6.

For FDG-PET, saliency maps exhibited spatially diffuse activation patterns distributed across temporoparietal association cortices and posterior cingulate regions. These patterns were bilaterally symmetric and extended across multiple contiguous slices in both coronal and axial views, indicating that model sensitivity was not restricted to a single focal region. Across subjects, FDG-PET attributions showed relatively consistent large-scale cortical involvement, with saliency extending along posterior and lateral cortical surfaces rather than being confined to deep medial structures.

In contrast, Tau-PET saliency maps demonstrated more spatially localized activation patterns, with dominant attribution concentrated in medial temporal lobe regions, including the hippocampus and entorhinal cortex. These activations were most prominent in sagittal and coronal views, where focal clusters were consistently observed across adjacent slices. Notably, saliency maps were confined to intracranial regions, with no systematic activation observed in skull or extracerebral structures.

Across both modalities, Grad-CAM maps displayed coherent anatomical structure and smooth spatial transitions rather than isolated voxel-level hotspots. This spatial continuity suggests that model predictions are driven by consistent regional patterns rather than noise or single-slice artifacts. Furthermore, the observed attribution patterns were stable across anatomical planes, supporting the robustness of the extracted saliency maps for visual assessment of modality-specific neuroanatomical emphasis.



(a) Coronal

(b) Axial

(c) Sagittal

Fig. 6: Comparative Grad-CAM saliency maps for FDG-PET and Tau-PET models across anatomical planes: (a-b) Coronal and axial views showing bilateral patterns, (c) sagittal view highlighting medial structures.

TABLE V: Mean Risk and Confidence Scores by Time-to-Conversion (TTC)

	FDG-PET			Tau-PET		
TTC Bin	**N**	**Risk Score**	**Conf. Score**	**N**	**Risk Score**	**Conf. Score**
$<$ 2 Years	128	0.68 ± 0.23	58.8 ± 29.8	23	0.53 ± 0.25	55.7 ± 33.5
2 – 4 Years	94	0.59 ± 0.21	41.5 ± 27.6	22	0.53 ± 0.26	48.6 ± 33.6
$>$ 4 Years	73	0.47 ± 0.22	41.5 ± 29.8	30	0.41 ± 0.18	37.3 ± 26.9

Values presented as Mean ± Standard Deviation.

E. Temporal Dynamics of Prediction

To evaluate the prognostic horizon of the models, we stratified both the predicted risk scores and the corresponding Confidence Scores for pMCI subjects based on their actual Time-to-Conversion (TTC). The results are summarized in Table V.

These temporal trends suggest distinct prognostic windows for metabolic versus protein biomarkers, a dynamic we explore further in the following discussion.

IV. Discussion

This work demonstrates that accurate prediction of progression from MCI to AD can be achieved from single-modality PET imaging while providing reliability and interpretability signals that are relevant for clinical decision support. In particular, the proposed framework couples competitive predictive performance with a Confidence Score that enables subject-level assessment of trust in each prediction, supporting reliability-aware risk stratification and helping clinicians distinguish confident cases from those that may require additional evaluation.

In addition, Grad-CAM-based spatial explainability provides an interpretable view of *where* the network focuses when estimating conversion risk, with attention patterns that are consistent with known disease-relevant neuroanatomy. Together, these components move the model beyond a point prediction and toward a clinically actionable tool that combines prognostic performance with transparent indicators of confidence and anatomical plausibility.

A. Generalization Across Modalities with Unified Preprocessing

A central finding of this study is that the Tau-PET model achieved mean classification accuracy comparable to the FDG-PET MCI model despite being trained on substantially fewer labeled subjects (200 vs. 627). The successful adaptation of FDG features to the Tau domain suggests that neurodegenerative patterns share a common structural basis across metabolic and protein imaging modalities. While global ranking performance (AUC) was lower for Tau-PET—likely due to sample size constraints—the binary classification accuracy remained equivalent, particularly in high-confidence regimes.

Importantly, the use of SynthStrip for skull-stripping enabled uniform preprocessing across both imaging modalities, eliminating potential domain shift arising from differential treatment of extracerebral signal. Although Tau-PET is known to exhibit off-target binding in skull and meningeal structures, SynthStrip's synthesis-driven training strategy [8] provided robust brain extraction without the cortical signal loss or masking inconsistencies observed with conventional methods. Comparative experiments demonstrated that this preprocessing approach substantially improved both predictive performance and the anatomical specificity of Grad-CAM attention maps, particularly for Tau-PET. This preprocessing standardization strengthens the validity of cross-modality transfer learning by ensuring that differences in model behavior reflect true biological signal rather than preprocessing artifacts.

B. Clinical Relevance and Trustworthiness

Beyond overall performance, the inclusion of uncertainty estimation provides an important layer of clinical interpretability. Stratifying predictions by Confidence Score revealed that high-confidence cases achieved high accuracy values across both imaging modalities, whereas low-confidence predictions approached chance-level performance. While FDG-PET reached approximately 90% accuracy in the high-confidence regime, Tau-PET achieved approximately 82%, likely reflecting the smaller training calibration set. Both modalities show progressive improvement across confidence bins, with FDG-PET exhibiting a steeper gradient from low (64.4%) to high (89.6%) confidence, while Tau-PET shows a more gradual progression from 65.0% to 82.1%. This suggests that while the model successfully filters out low-quality predictions (low-confidence bin), the scarcity of Tau data limits its ability to fully refine calibration at the extreme high-confidence end. Overall, this behavior suggests that uncertainty estimates capture meaningful information about prediction reliability rather than reflecting random noise.

Such stratification supports a reliability-aware decision support paradigm in which model outputs are interpreted in conjunction with their associated confidence. High-confidence predictions may be used to support clinical planning and follow-up, while low-confidence cases can be flagged for additional diagnostic evaluation, such as CSF biomarker analysis or complementary imaging. Importantly, this framework is intended to assist, rather than replace, clinical judgment.

The biological plausibility of model predictions further strengthens confidence in the proposed approach. Predicted conversion risk correlated negatively with cognitive performance (MMSE and MoCA) and positively with APOE4 carrier status, aligning with established clinical and genetic risk factors. Correlations were generally stronger for FDG-PET than for Tau-PET, particularly with cognitive measures, consistent with the interpretation that FDG-PET reflects current synaptic dysfunction and metabolic impairment more directly linked to contemporaneous cognitive status.

Grad-CAM analyses revealed anatomically meaningful attention patterns, with FDG-PET highlighting temporoparietal and posterior cingulate regions and Tau-PET focusing on medial temporal lobe structures associated with early tau pathology. For Tau-PET, this pattern aligns with Braak stages I–II and confirms that SynthStrip preprocessing enabled the model to focus on biologically meaningful cortical tau distributions without interference from extracerebral uptake. For FDG-PET, the highlighted regions are consistent with established biomarkers showing early hypometabolism in prodromal Alzheimer's disease.

Together, these findings suggest that the model relies on disease-relevant signals rather than spurious artifacts.

These findings support the interpretation of uncertainty estimates as complementary decision-support information, enabling differentiation between stable and unstable predictions without altering the underlying classification threshold.

C. Temporal Sensitivity and Biomarker Staging

Our Time-to-Conversion analysis reveals distinct temporal behaviors for metabolic and protein biomarkers. The FDG-PET model exhibits a strong temporal gradient, with higher predicted risk scores for subjects converting within two years and progressively lower scores for more distal converters. This pattern is further reflected in the associated Confidence Scores, which are highest in the <2 year window and decrease substantially for longer time-to-conversion intervals. Together, these results indicate that FDG-PET provides both stronger risk signals and higher predictive certainty when conversion is imminent, consistent with synaptic dysfunction as a downstream process occurring closer to clinical onset [15].

In contrast, the Tau-PET model demonstrates relative stability in mean risk scores across the <2 and 2–4 year intervals, while Confidence Scores show a gradual decline with increasing time-to-conversion. This suggests that Tau-PET captures an earlier, trait-like pathological signal that is present years before diagnosis, but with reduced temporal specificity for distant converters. The lower overall risk scores and confidence values observed for Tau-PET likely reflect calibration effects arising from the smaller training cohort ($N = 200$) rather than an absence of biologically meaningful signal. As larger tau PET datasets become available, this modality may offer improved early-window prognostic performance with greater temporal resolution.

Notably, for intermediate converters (2–4 years), Tau-PET exhibits higher mean Confidence Scores than FDG-PET despite comparable risk estimates, suggesting more stable uncertainty in this temporal window.

D. Limitations

Several limitations should be acknowledged.

First, the primary bottleneck for the Tau-PET model is the data scarcity within the ADNI cohort. While the FDG model benefited from a robust training set ($N = 627$), the strict inclusion criteria required for the conversion study (valid baseline Tau-PET + sufficient longitudinal follow-up) reduced the usable Tau cohort to 200 subjects. Although transfer learning successfully mitigated this by leveraging FDG-learned features, the reduced sample size inherently limits the model's exposure to the full spectrum of tau pathology.

Second, although ADNI is a large, multi-site study, it may not fully capture the heterogeneity encountered in routine clinical populations, underscoring the need for external validation on independent cohorts such as AIBL or OASIS.

We acknowledge that deep learning on small cohorts ($N = 200$) carries overfitting risks compared to ROI-based tabular models. However, we prioritized a voxel-wise CNN approach to enable (1) the transfer of fine-grained spatial features from the larger FDG dataset, and (2) the generation of granular Grad-CAM distinct from predefined anatomical atlases, which is crucial for the explainability objectives of this study.

Finally, resampling volumes to $100 \times 100 \times 90$ voxels entails a trade-off between computational efficiency and spatial resolution. While explainability analyses suggest that key anatomical regions were preserved, future work could explore higher-resolution inputs or multiscale architectures to capture finer pathological details.

E. Future Work

Future work could focus on strengthening the Tau-PET setting by augmenting the cohort with additional datasets and external populations, thereby increasing sample diversity and improving the robustness of the learned representations. In particular, incorporating harmonization strategies and external validation protocols would help assess generalizability beyond the ADNI cohort.

In addition, although Grad-CAM was employed for qualitative interpretability, a promising direction is to systematically relate spatial attributions to prediction uncertainty. Analyzing Grad-CAM patterns for low-confidence and misclassified subjects may provide insight into the sources of model ambiguity, inform refinements to the training strategy or network architecture, and help identify recurring imaging characteristics associated with uncertain or unstable predictions.

V. Conclusion

We presented a deep learning framework for predicting MCI-to-AD progression that bridges the data availability gap between established FDG-PET and emerging Tau-PET imaging through hierarchical cross-modality transfer learning. By integrating uncertainty estimation and explainability, the proposed approach moves beyond point predictions and provides reliability-aware, biologically interpretable outputs. These characteristics are essential for the development of trustworthy AI systems and support the potential clinical utility of deep learning models in neurodegenerative disease prognosis.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Data and Code Availability

The source code for the data preprocessing, model training, and evaluation pipelines developed in this study is available at: `https://github.com/duartemoura/sMCI-vs-pMCI-Alzheimer`.

The neuroimaging data used in this work were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As per ADNI data use agreements, the raw imaging data cannot be publicly redistributed; however, they are available to authorized researchers upon request and approval by the ADNI Data Sharing and Publications Committee.

References

[1] M. Fernandez-Garcia, et al., "Improving confidence in long-term deep learning prediction of progression from MCI to AD using ^{18}F-FDG-PET," Master's Thesis, Universidad Carlos III de Madrid, 2024.

[2] R. R. Selvaraju, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proc. ICCV*, 2017.

[3] T. Jo, et al., "Deep learning detection of informative features in tau PET for Alzheimer's disease classification," *BMC Bioinformatics*, 2020.

[4] R. Santangelo, et al., "CSF p-tau/Aβ42 ratio and brain FDG-PET may reliably detect MCI 'imminent' converters to AD," *Eur. J. Nucl. Med. Mol. Imaging*, 2020.

[5] P. Zhou, et al., "Deep-Learning Radiomics for Discrimination Conversion of Alzheimer's Disease," *Front. Aging Neurosci.*, 2021.

[6] L. A. De Santi, et al., "An Explainable Convolutional Neural Network for the Early Diagnosis of Alzheimer's Disease," *J. Digit. Imaging*, 2022.

[7] ADNI, "PET Pre-processing protocols," [Online]. Available: http://adni.loni.usc.edu.

[8] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "SynthStrip: Skull-Stripping for Any Brain Image," *NeuroImage*, vol. 260, p. 119474, 2022, doi: 10.1016/j.neuroimage.2022.119474.

[9] M. Marquié, et al., "Lessons learned about [^{18}F]-AV-1451 off-target binding from an autopsy-confirmed Parkinson's case," *Acta Neuropathol. Commun.*, vol. 5, no. 1, p. 75, Oct. 2017, doi: 10.1186/s40478-017-0482-0.

[10] S. Flores, et al., "Investigating Tau and Amyloid Tracer Skull Binding in Studies of Alzheimer Disease," *J. Nucl. Med.*, vol. 64, no. 2, pp. 287–293, Feb. 2023, doi: 10.2967/jnumed.122.263948.

[11] F. J. López-González, et al., "Impact of spill-in counts from off-target regions on [^{18}F]Flortaucipir PET quantification," *NeuroImage*, vol. 259, p. 119396, Oct. 2022, doi: 10.1016/j.neuroimage.2022.119396.

[12] M. R. Scott, et al., "Contribution of extracerebral tracer retention and partial volume effects to sex differences in Flortaucipir-PET signal," *J. Cereb. Blood Flow Metab.*, vol. 44, no. 1, pp. 131–141, Jan. 2024, doi: 10.1177/0271678X231196978.

[13] D. Alvarez-Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. 32nd Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7786–7795.

[14] A. Bhatt, P. Ravikumar, and R. Srikant, "On the limits of self-explainable models," *arXiv preprint arXiv:2410.02331*, 2024.

[15] C. R. Jack Jr, et al., "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[16] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[17] E. H. Corder, et al., "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.