INDIANA UNIVERSITY
School of Medicine

AAIC 2026 · London · 12 to 15 July

Uncertainty-aware tau-PET classification of Alzheimer's disease: calibration, selective prediction, and clinical utility

Taeho Jo, Eun Hye Lee, Kwangsik Nho, Andrew J. Saykin, for the Alzheimer's Disease Neuroimaging Initiative
Indiana University School of Medicine, Indianapolis, IN, USA tjo@iu.edu

Background

- Tau PET imaging quantifies the neurofibrillary pathology that most closely tracks cognitive decline in AD, and accurate individual-level tools are needed for early detection and selection for anti-amyloid therapies such as lecanemab and donanemab.
- Deep learning classifiers on regional tau SUVR perform well but return only a point estimate. Per-patient reliability is not quantified, and softmax scores are often interpreted as probabilities without calibration.
- TRIPOD+AI recommends evaluating prediction models on three axes: discrimination, calibration, and clinical utility. Most neuroimaging classifiers report discrimination alone.
- Monte Carlo dropout approximates a Bayesian posterior, yielding a per-patient uncertainty that can drive selective prediction.

Methods

Cohort and features

- 691 ADNI participants (575 cognitively normal, 116 AD dementia). 114 regional tau PET SUVRs used as input, with no feature selection or dimensionality reduction.

Model and uncertainty

- Feature-Tokenizer Transformer (FT-Transformer) with Monte Carlo dropout. Each participant is scored over 50 stochastic passes; the mean is the predicted probability of AD and the standard deviation is the per-participant uncertainty.

Training

- AdamW with cosine learning-rate schedule, cross-entropy loss, up to 200 epochs with early stopping (patience 25), and 30 trials of random hyperparameter search.

Validation and leakage control

- Nested cross-validation (5 outer, 3 inner folds). Imputation and standardization are fit inside training folds only; out-of-fold pooling; bootstrap 95% CIs (2,000 resamples).

Evaluation

- Discrimination, calibration (ECE, Brier, slope and intercept), the uncertainty-error relationship, selective prediction (AURC), and decision curve analysis. Reporting follows TRIPOD+AI and CLAIM.

Table 1. Baseline characteristics

Characteristic	CN (n=575)	ADD (n=116)
Age, years	71.3 ± 7.9	76.0 ± 8.9
Female	61.9%	42.2%
Education, years	16.7 ± 2.3	15.6 ± 2.5
APOE ε4 carrier	32.9%	63.8%

Discrimination and uncertainty

- The FT-Transformer achieved AUROC 0.903 (95% CI 0.868-0.934) and accuracy 90.7%. Precision 0.770, recall 0.646, and F1 0.697 reflect class imbalance (116 AD vs 575 CN).
- Predictive uncertainty tracked classification error. The highest-uncertainty quartile had 13.0x the error rate of the lowest (Q4 22.7% vs Q1 1.7%; 95% CI 4.85-75.0, Fisher P < 0.001).
- Uncertainty was higher among misclassified than correct cases (Mann-Whitney U = 30,952, P = 4.1e-13), discriminating errors with AUROC 0.771 (Cramer's V 0.311).

0.903
AUROC (95% CI 0.868-0.934)

13.0x
Q4 vs Q1 error-rate ratio

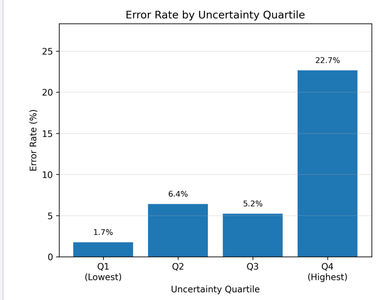

Figure 1A. Error rate by uncertainty quartile

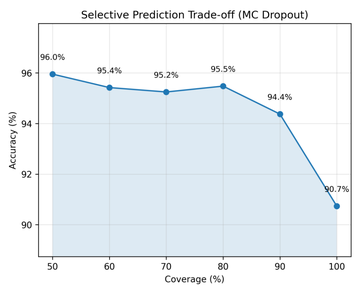

Figure 1B. Selective-prediction accuracy versus coverage

Selective prediction

- Deferring the most uncertain cases raised accuracy on the retained set from 90.7% at full coverage to 95.5% at 80% (138 deferred) and 96.0% at 50% (345 deferred).
- Area under the risk-coverage curve was 0.035, so errors concentrate in the deferred fraction. The uncertainty threshold is explicit and tunable per site.

Calibration

- Probabilities were well calibrated without any post-hoc correction: ECE 0.028, Brier 0.070.
- Calibration slope 0.911 and intercept -0.216 indicate only mild overconfidence at the extremes of the probability range.

Table 2. Calibration metrics

Calibration metric	Value
Expected calibration error (ECE)	0.028
Brier score	0.070
Calibration slope	0.911
Calibration intercept	-0.216

Table 3. Selective prediction by uncertainty coverage

Coverage	Threshold	Accuracy	Errors
100%	0.192	90.7%	64
90%	0.086	94.4%	35
80%	0.058	95.5%	25
70%	0.042	95.2%	23
60%	0.035	95.4%	19
50%	0.030	96.0%	14

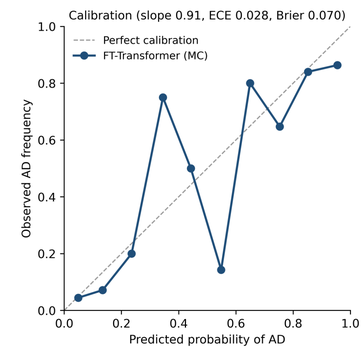

Figure 2. Reliability diagram

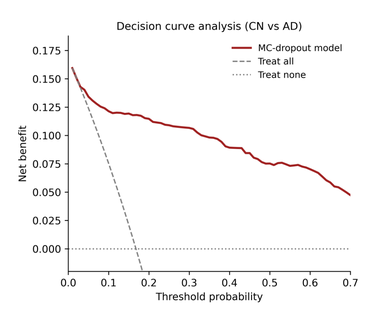

Figure 3. Decision curve versus treat-all and treat-none.

Clinical utility

- Decision curve analysis showed net benefit above both treat-all and treat-none across threshold probabilities 0.03 to 0.70, spanning screening to confirmatory decisions.
- Within this range the model captured true positives without an offsetting false-positive cost, consistent with a screening or triage instrument.

Conclusions

- An uncertainty-aware tau PET classifier was well calibrated, ranked its own errors, and added net benefit across clinically plausible thresholds.
- A workflow auto-reporting roughly four of five reads at 95.5% accuracy and routing the rest to a neuroradiologist could reduce reading burden while maintaining quality.
- Limitations. Single cohort (ADNI), no external validation, predominantly European-ancestry and highly educated, and class imbalance (recall 0.646).
- Next. External validation across scanners and tracers, multimodal data, and prospective clinical evaluation. Data from ADNI (adni.loni.usc.edu).

Resources and funding

- Funding Alzheimer's Association AARG-22-974053 and NIH U01 AG068057, R01 AG092591, U19 AG074879, U01 AG024904, P30 AG072976, U01 AG072177