# An Adaptive Machine Learning Triage Framework for Predicting Alzheimer's Disease Progression

**Richard Hou**[1] RICHARD.HOU@EMORY.EDU
**Shengpu Tang**[2,†] SHENGPU.TANG@EMORY.EDU
**Wei Jin**[2,†] WEI.JIN@EMORY.EDU

[1] *Department of Biology, Emory University, USA*
[2] *Department of Computer Science, Emory University, USA*
[†] *Equal contribution as senior authors.*

## Abstract

Accurate predictions of conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD) can enable effective personalized therapy. While cognitive tests and clinical data are routinely collected, they lack the predictive power of PET scans and CSF biomarker analysis, which are prohibitively expensive to obtain for every patient. To address this cost-accuracy dilemma, we design a two-stage machine learning framework that selectively obtains advanced, costly features based on their predicted "value of information". We apply our framework to predict AD progression for MCI patients using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Our framework reduces the need for advanced testing by 20% while achieving a test AUROC of 0.929, comparable to the model that uses both basic and advanced features (AUROC=0.915, $p$=0.1010). We also provide an example interpretability analysis showing how one may explain the triage decision. Our work presents an interpretable, data-driven framework that optimizes AD diagnostic pathways and balances accuracy with cost, representing a step towards making early, reliable AD prediction more accessible in real-world practice. Future work should consider multiple categories of advanced features and larger-scale validation.

**Keywords:** disease progression prediction, Alzheimer's disease, uncertainty estimation

**Data and Code Availability** We used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] (adni.loni.usc.edu). The data is publicly available and can be requested on the ADNI website. Code for our experiments is available online.[2]

**Institutional Review Board (IRB)** This work does not require IRB approval.

## 1. Introduction

Alzheimer's disease (AD), a major cause of dementia, is a neurodegenerative disorder marked by memory loss, cognitive decline, behavioral alterations, and diminished functional capabilities (Vaz and Silvestre, 2020). Over 40 million people worldwide currently suffer from dementia (Golde, 2022). A critical window for intervention lies in the stage of mild cognitive impairment (MCI), a transitional state where individuals exhibit cognitive deficits but maintain functional independence (Kelley and Petersen, 2007). However, MCI is a heterogeneous condition; while many patients progress to AD, a significant portion remains stable or even reverts to normal cognition (Canevelli et al., 2016). Thus, identifying patients with MCI who are more likely to progress to AD is crucial to enable effective, personalized therapy (Golde et al., 2018; Li et al., 2021).

Machine learning (ML) offers a promising solution to this problem by integrating complex, multimodal data to identify patterns that differentiate between progressive and stable MCI (Grueso and Viejo-Sobera, 2021). However, the predictive power of ML models depends on the quality of the input features (Domingos, 2012). Biomarkers derived from positron emission tomography (PET) scans or cerebrospinal

1. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.
2. https://github.com/chardhou-cpu/Triage-Framework-AD

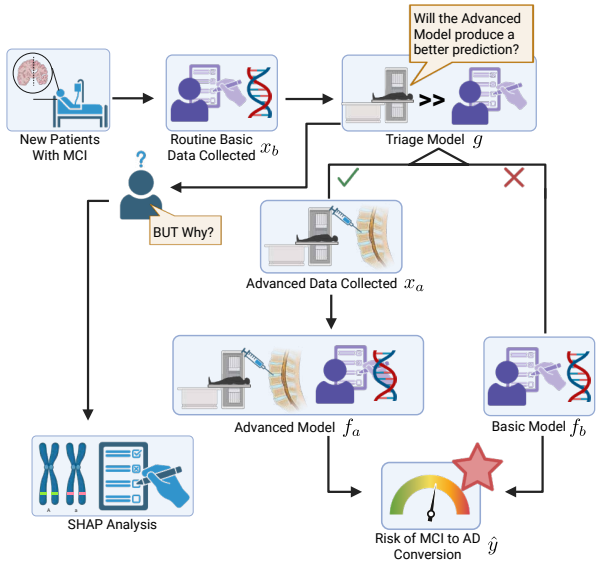


Figure 1: Our proposed two-stage framework for AD progression prediction.

fluid (CSF) analysis are highly informative but can be too expensive and invasive for routine screening (Spasov et al., 2019). In contrast, low-cost features such as cognitive tests are more accessible but often lack diagnostic precision (Chen et al., 2022; Mitchell, 2015; Han et al., 2017). This trade-off between cost and accuracy limits reliable AD detection to specialized centers and impedes the development of scalable risk stratification approaches.

To address this challenge, we propose an **adaptive, two-stage** ML framework that mirrors how clinicians reason about initial screening and targeted follow-up testing (Figure 1). Our framework consists of three models: the **Basic Model**, the **Advanced Model**, and the **Triage Model**. The Basic Model leverages only low-cost, widely accessible clinical data and can be applied to all patients. The Advanced Model combines routine clinical data with biomarkers and PET scan data, and thus is only applicable to patients who have undergone these advanced testing. At the core of our framework is the Triage Model, which determines if a patient should undergo advanced testing. This decision is based on whether such testing is expected to *improve diagnostic certainty*, ensuring that costly resources are used only where they are most likely to have an impact. In addition, the Triage Model is designed with interpretability in mind, providing clinicians with clear explanations of why escalation is recommended.

Our contributions are threefold: (1) We propose a cascading model that selectively allocates advanced tests for patients, which effectively reduces average diagnostic cost while preserving accuracy. (2) We develop an interpretable triage strategy that explicitly estimates when escalation is warranted, identifying cases in which advanced features are expected to provide meaningful diagnostic benefit. (3) We validate our framework on the ADNI dataset, demonstrating comparable AUROC to full-data models while lowering feature acquisition costs by up to 20%. This work demonstrates how adaptive prediction can align machine learning with real-world clinical priorities, thereby advancing the development of cost-sensitive and interpretable systems for AD prediction.

## 2. Methods

### 2.1. Method Overview

Our framework consists of three model components (Figure 1): the Basic Model $f_b : \mathbb{R}^{d_b} \to [0,1]$, the Advanced Model $f_a : \mathbb{R}^{d_b+d_a} \to [0,1]$, and the Triage Model $g : \mathbb{R}^{d_b} \to [0,1]$. We use $d_b$ and $d_a$ to denote the dimensionality of the basic and advanced feature sets, respectively, with $x_b$ and $x_a$ denoting the corresponding feature vectors. We refer to the output $g(x)$ of the Triage Model as the escalation score. The final prediction $\hat{y}$ is determined as:

$$\hat{y} = \begin{cases} f_b(x_b) & \text{if } g(x_b) \le \tau, \\ f_a(x_b, x_a) & \text{if } g(x_b) > \tau, \end{cases}$$

where the escalation threshold $\tau \in [0,1]$ is a hyperparameter that trades off coverage (fraction of patients handled by the Basic Model) and selective risk (error rate when not escalating). In other words, the Basic Model $f_b$ is used by default; if escalation is deemed necessary according to the Triage Model, the Advanced Model $f_a$ is used. Let $p_b$ and $p_a$ denote the predicted probabilities for the positive class from the Basic and Advanced Models, respectively. We define the certainty of the model prediction as the absolute distance between the predicted probability and 0.5, $c = |p - 0.5|$. To train the Triage Model $g$, we use a binary supervision signal: $z = \mathbf{1}[c_a - c_b > \delta]$, where $\delta > 0$ is a margin parameter that specifies how much more certain the Advanced Model must be compared to the Basic Model to justify escalation. In this way, $z = 1$ indicates that escalation provides a meaningful gain in certainty, while $z = 0$ indicates that the Basic Model is likely already sufficiently certain. Thus,

the Triage Model $g(x_b)$ is trained to anticipate when escalation to the Advanced Model is worthwhile.

**Interpretability of the Escalation Decision.** To make the escalation decisions transparent, we use SHAP (Lundberg and Lee, 2017), a widely adopted method for feature attribution, to the Triage Model. For each patient, SHAP assigns a contribution value $\phi_j$ to feature $j$, representing the direction and magnitude of the feature's impact. Positive $\phi_j > 0$ indicates that larger values of a feature increase the likelihood of escalation, and vice versa. In practice, we present the features with the strongest contributions together with their raw clinical values (e.g., MMSE score = 27/30 with $\phi = +0.81$).

### 2.2. Implementation Details

**Data Source and Cohort**. All data are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI is to test whether serial MRI, PET, biological markers, and clinical assessments could be combined to measure the progression of MCI and early AD. Our final analysis include 1,142 participants who had been diagnosed with MCI.

**Feature Sets and Prediction Target.** The Basic Model is trained using a basic feature set ($d_b = 9$), which includes demographics (age, gender, race, education), APOE genetic status, and cognitive scores (MMSE, ADAS-11, Global CDR). The Advanced Model uses both the basic features and an advanced feature set ($d_a = 329$) containing biomarker data from CSF analysis (Aβ_42, t-tau, and p-tau levels) and PET imaging (amyloid status, hippocampal volume, etc). Advanced features are available for 551 participants. For both the Basic and Advanced Models, the prediction target is a binary indicator of whether a patient had converted from MCI to AD over a two-year period. Finally, the Triage Model uses the same feature set as the Basic Model but is trained on the binary label defined in Section 2.1.

**Data Partitioning.** We carefully partition the available data to ensure robust evaluation. A held-out test set (n=100) is randomly sampled from the cohort (n=551) for which both basic and advanced features are available. The Basic Model is trained on the remaining cohort (n=1,042) who have basic features. The Advanced and Triage Models are trained on the remaining cohort (n=451) who have both basic and advanced features.

**Model Implementation.** All models are built using scikit-learn pipelines that included preprocessing steps such as standardization of numerical features and one-hot encoding for categorical features, with 5-fold cross validation for hyperparameter and model selection. The final models are L2-regularized logistic regression classifiers for both the Basic Model and Advanced Model, and a support vector machine classifier with RBF kernel for the Triage Model, The final, optimized hyperparameters for each model are detailed in Appendix A. To prevent data leakage in training the Triage Model, its labels are generated using 5-fold cross-validation predictions of the Basic and Advanced Models.

**Evaluation and Baselines.** We evaluate our framework on the held-out test set using standard binary classification metrics including AUROC, AUPRC, accuracy, precision, and recall. The escalation threshold for the Triage Model is selected by analyzing a risk-coverage curve on the training set. We compare our approach against several baselines with different heuristics, including the Basic Model, the Advanced Model, random escalation, and escalation based on the Basic Model's output probability or uncertainty. We also display a Cost-AUROC curve to illustrate the trade-off between diagnostic performance and the expected financial cost of testing.

## 3. Results

**Performance of Base Models.** The Advanced Model achieves an AUROC of 0.915 (95% CI: 0.847-0.968) on the held-out test set, outperforming the Basic Model with an AUROC of 0.791 (95% CI: 0.686-0.879) (Table 1). While the basic features already achieves nontrivial performance, the advanced features provide a significant boost in predictive power.

**Triage Threshold Selection.** We set the certainty gain margin $\delta$ to be 0.2, which created a reasonably balanced supervision signal for training the Triage Model while capturing a "meaningful" increase in prediction certainty. After hyperparameter tuning, the final Triage Model achieves a cross-validation AUROC of 0.77 in predicting the binary label representing the expected certainty gain. Based on the risk-coverage curve (Figure 2) that illustrates the trade-off between coverage (the percentage of patients handled by the Basic Model) and risk (the error rate for the non-escalated group), we observe an inflection point at 19% coverage, where the error rate drops sharply to 8.0%. This corresponds to an escalation threshold of $\tau = 0.05$.

**Final Performance Comparison.** The performance of our proposed framework is summarized in

Table 1: Performance comparison with 95% bootstrapped confidence intervals on the test set.

| Model | AUROC (%) | AUPRC (%) | Accuracy (%) | Recall (%) | Precision (%) |
|---|---|---|---|---|---|
| Basic Model Alone | 79.1 (68.6 - 87.9) | 67.3 (48.8 - 81.8) | 72.0 (63.0 - 80.0) | 58.1 (40.0 - 74.1) | 54.5 (37.0 - 72.4) |
| Advanced Model Alone | 91.5 (84.7 - 96.8) | 85.2 (72.1 - 94.2) | 82.0 (74.0 - 89.0) | 90.3 (78.1 - 100.0) | 65.1 (50.0 - 79.1) |
| Our Triage Model (Thresh=0.05) | 92.8 (86.1 - 97.6) | 89.2 (78.9 - 95.9) | 83.0 (75.0 - 90.0) | 90.3 (78.1 - 100.0) | 66.7 (51.4 - 81.1) |
| Baseline: Random 80% Escalate | 89.0 (81.4 - 94.5) | 79.3 (64.0 - 90.7) | 80.0 (72.0 - 87.0) | 80.6 (66.7 - 93.3) | 64.1 (48.5 - 79.2) |
| Baseline: Escalate 80% Highest Prob | 91.7 (84.5 - 96.9) | 85.5 (72.4 - 94.4) | 83.0 (75.0 - 90.0) | 90.3 (78.1 - 100.0) | 66.7 (51.4 - 81.1) |
| Baseline: Escalate 80% Most Uncertain | 92.6 (85.7 - 97.4) | 88.7 (78.1 - 95.5) | 83.0 (75.0 - 90.0) | 90.3 (78.1 - 100.0) | 66.7 (51.4 - 81.1) |

Table 1, along with baselines that have the same escalation rate (80%). Our proposed triage framework achieves an AUROC of 0.928 (95% CI: 0.861-0.976) and an AUPRC of 0.892 (95% CI: 0.789-0.959), outperforming the random escalation baseline and escalation based on Basic Model's probability predictions, and marginally surpassing escalation based on Basic Model's most uncertain cases. Notably, our triage framework achieves comparable performance to using the Advanced Model alone. Based on the PR curves (Figure 6), our proposed framework can correctly identify nearly 60% of positive cases (recall) while maintaining 100% precision and 0 false positives, whereas all other baselines can only achieve a recall of $< 30\%$ under the same condition.

**Interpreting Triage Decisions.** We select two representative patients to illustrate how the Triage Model generates interpretable decisions. For the first case (Figure 3-top), the Triage Model assigned a high escalation score of 0.50, indicating a recommendation *for* escalation to the Advanced Model. The most influential features include a high normalized ADAS-11 total score (0.43), a high normalized MMSE score (0.56), and the absence of the APOE4 allele. However, lower ADAS-11 and higher MMSE scores reflect good cognitive performance clinically (Cipolotti and Warrington, 1995), and the absence of APOE4 removes a major genetic risk factor for AD (Corder et al., 1993). In this specific case, the high ADAS-11 score contradicts the high MMSE score and the absence of APOE4 allele. This conflicting profile justifies the need for an Advanced Test to confirm the diagnosis. For the second case (Figure 3-bottom), the Triage Model assigned a low escalation score of 0.04, indicating a recommendation *against* escalation, with the most influential features being a low normalized ADAS-11 total score ($-0.85$), a high normalized MMSE score (1.12), and the absence of the APOE4 allele. Together, these characteristics form a coherent profile consistent with "Stable MCI", justifying the decision not to escalate.

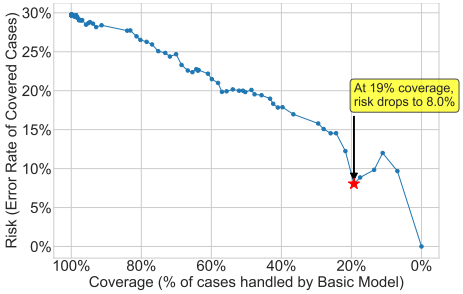


Figure 2: Risk-coverage curve and the selected threshold for the Triage Model.

**Cost-Effectiveness Analysis.** Figure 4 shows the relationship between discriminative performance and the expected financial cost per 100 patients. For comparison, we also include the performance of the Basic Model alone (AUROC=0.79 at zero cost) and the Advanced Model alone (AUROC=0.92 at a maximum cost of $400,000). In general, AUROC improves as the number of escalations increases. Our selected operating point escalates approximately 80% of patients, saving approximately $80,000 per 100 patients (Alzheimer's Association, 2025) compared to the Advanced Model while achieving a higher AUROC.

## 4. Discussion & Conclusion

In this work, we propose an adaptive two-stage ML framework for cost-effective prediction of Alzheimer's disease progression. Our framework integrates three models: a Basic Model trained on low-cost features, an Advanced Model that leverages additional costly features, and a Triage Model that selectively escalates patients when the additional features are expected to reduce model uncertainty. Empirically, our framework achieves performance comparable to the Advanced Model while potentially reducing resources required for obtaining expensive features. SHAP-based analysis further demonstrates that the triage decisions are clinically interpretable, with cognitive and genetic factors emerging as key drivers.

The potential real-world impact of our proposed framework is significant on both patients and the

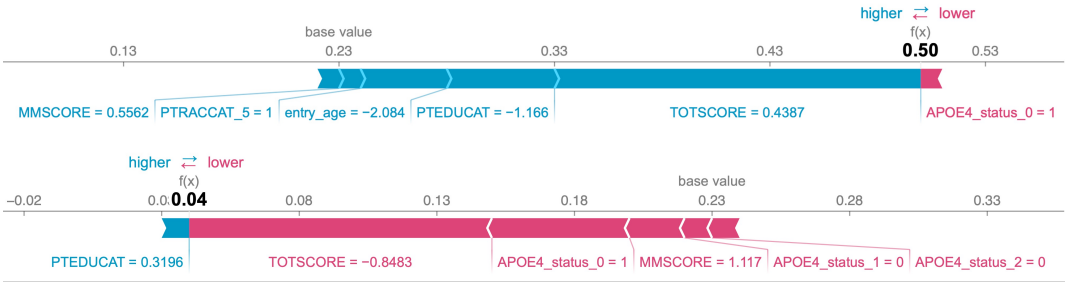


Figure 3: SHAP force plots explaining the Triage Model's predictions recommending escalation (top) and recommending no escalation (bottom). TOTSCORE - normalized ADAS-11 total score; MMSCORE - normalized MMSE score; APOE4_status - APOE4 carrier status.

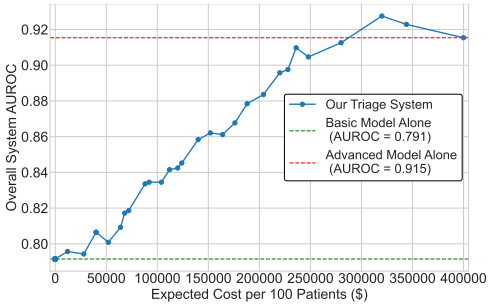


Figure 4: Cost-AUROC trade-off curve.

healthcare system. By accurately identifying 20% of individuals with MCI who do not require escalation, our framework directly reduces patient burden. This could translate to 2.4 million individuals in the U.S. potentially avoiding invasive and costly procedures. The financial implications are equally significant, representing a potential societal cost saving of over $9.6 billion (Alzheimer's Association, 2025). Therefore, our work demonstrates a powerful pathway to not only conserve clinical resources but, more importantly, to protect patients from unnecessary interventions.

Although our triage framework achieved statistically comparable performance to the uncertainty-based escalation baseline (Table 3), this does not mean that the Triage Model is redundant. Rather, it offers three distinct advantages:

1. Directness: The Triage Model is explicitly trained to predict whether performing an advanced test will yield a meaningful gain in diagnostic certainty. In contrast, the Basic Model's uncertainty only serves as a proxy; while correlated, it does not necessarily indicate that an advanced test will reduce uncertainty.
2. Interpretability: The dedicated Triage Model enables explanation of escalation decisions using techniques such as SHAP. These explanations (e.g., Figure 3) are more transparent and informative than a generic statement that "the Basic Model is uncertain".
3. Adaptability: While our paper considers a two-stage process with a binary escalation decision, the Triage Model naturally generalizes to a multi-class setting that could select among multiple advanced tests. Our approach thus lays important groundwork for devising more sophisticated, multi-step diagnostic protocols.

**Limitations and Future Directions.** Due to the small dataset size, the training split was reused for Triage Model training and escalation threshold selection; future work with larger cohorts should use separate validation sets to ensure robust generalization. While we have defined the advanced feature set to include multiple data sources (CSF biomarkers and PET imaging), future work could consider separating these modalities to develop more granular, multi-step diagnostic pathways using techniques such as reinforcement learning (Tang, 2024). Our framework uses static, cross-sectional data from a patient's baseline visit. Incorporating longitudinal data (e.g., time-series analysis of cognitive scores) would be a valuable extension and likely improve timeliness of diagnosis and actionability of the testing decisions (Tang, 2025). The interpretability of our Triage Model further enables clinician-in-the-loop decision making (Tang et al., 2020), where model explanations could support shared decision processes between human experts and AI. Future user studies involving clinicians could evaluate how such explanations affect trust, usability, and real-world adoption. Finally, our framework could be applied to other clinical tasks beyond predicting two-year AD progression. Future work could also compare our triage model against frontier approaches for producing escalation decisions, including large language models and AI agents (Xu et al., 2025).

## Acknowledgments

We thank the anonymous reviewers of ML4H 2025 for their valuable feedback. All content and code were verified by the authors, who take full responsibility for the results. This work was partially supported by the U.S. National Science Foundation under Award Numbers 2504088 and 2437345 (to W.J.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

Alzheimer's Association. 2025 Alzheimer's Disease Facts and Figures. Technical report, 2025.

Marco Canevelli, Giulia Grande, Eleonora Lacorte, Elisa Quarchioni, Matteo Cesari, Claudio Mariani, Giuseppe Bruno, and Nicola Vanacore. Spontaneous Reversion of Mild Cognitive Impairment to Normal Cognition: A Systematic Review of Literature and Meta-Analysis. *Journal of the American Medical Directors Association*, 17(10):943–948, October 2016. ISSN 1538-9375. doi: 10.1016/j.jamda.2016.06.020.

Yanru Chen, Xiaoling Qian, Yuanyuan Zhang, Wenli Su, Yanan Huang, Xinyu Wang, Xiaoli Chen, Enhan Zhao, Lin Han, and Yuxia Ma. Prediction Models for Conversion From Mild Cognitive Impairment to Alzheimer's Disease: A Systematic Review and Meta-Analysis. *Frontiers in Aging Neuroscience*, 14, April 2022. ISSN 1663-4365. doi: 10.3389/fnagi.2022.840386. Publisher: Frontiers.

L Cipolotti and E K Warrington. Neuropsychological assessment. *Journal of Neurology, Neurosurgery, and Psychiatry*, 58(6):655–664, June 1995. ISSN 0022-3050. doi: 10.1136/jnnp.58.6.655.

E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance. Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science*, 261(5123):921–923, August 1993. doi: 10.1126/science.8346443. Publisher: American Association for the Advancement of Science.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55 (10):78–87, October 2012. ISSN 0001-0782, 1557-7317. doi: 10.1145/2347736.2347755.

Todd E. Golde. Alzheimer's disease - the journey of a healthy brain into organ failure. *Molecular Neurodegeneration*, 17(1):18, March 2022. ISSN 1750-1326. doi: 10.1186/s13024-022-00523-1.

Todd E. Golde, Steven T. DeKosky, and Douglas Galasko. Alzheimer's disease: The right drug, the right time. *Science*, 362(6420):1250–1251, December 2018. ISSN 1095-9203. doi: 10.1126/science.aau0437.

Sergio Grueso and Raquel Viejo-Sobera. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Alzheimer's Research & Therapy*, 13(1):1–29, December 2021. ISSN 1758-9193. doi: 10.1186/s13195-021-00900-w. Publisher: BioMed Central.

Xiaoxia Han, Yilong Zhang, and Yongzhao Shao. Application of Concordance Probability Estimate to Predict Conversion from Mild Cognitive Impairment to Alzheimer's Disease. *Biostatistics & Epidemiology*, 1(1):105–118, 2017. ISSN 2470-9360. doi: 10.1080/24709360.2017.1342187.

Brendan J. Kelley and Ronald C. Petersen. Alzheimer's Disease and Mild Cognitive Impairment. *Neurologic Clinics*, 25(3):577–v, August 2007. ISSN 0733-8619. doi: 10.1016/j.ncl.2007.03.008.

Hai-Tao Li, Shao-Xun Yuan, Jian-Sheng Wu, Yu Gu, and Xiao Sun. Predicting Conversion from MCI to AD Combining Multi-Modality Data and Based on Molecular Subtype. *Brain Sciences*, 11(6):674, May 2021. ISSN 2076-3425. doi: 10.3390/brainsci11060674.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Alex J. Mitchell. Can the MMSE help clinicians predict progression from mild cognitive impairment to dementia?: Commentary on... Cochrane Corner. *BJPsych Advances*, 21(6):363–366, November 2015. ISSN 2056-4678, 2056-4686. doi: 10.1192/apt.21.6.363.

Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Liò, and Nicola Toschi. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. *NeuroImage*, 189:276–287, April 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2019.01.031.

Shengpu Tang. *Towards Clinically Applicable Reinforcement Learning*. PhD thesis, University of Michigan, 2024.

Shengpu Tang. Transforming healthcare decision making using artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (27):28729–28729, 2025.

Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In *International Conference on Machine Learning*, pages 9387–9396. PMLR, 2020.

Miguel Vaz and Samuel Silvestre. Alzheimer's disease: Recent treatment strategies. *European Journal of Pharmacology*, 887:173554, November 2020. ISSN 1879-0712. doi: 10.1016/j.ejphar.2020.173554.

Gelei Xu, Xueyang Li, Yixiong Chen, Yuying Duan, Shuqing Wu, Alexander Yu, Ching-Hao Chiu, Juntong Ni, Ningzhi Tang, Toby Jia-Jun Li, Alan Yuille, Wei Jin, and Yiyu Shi. A Comprehensive Survey of Agentic AI in Healthcare, November 2025.

Gelei Xu, Yuying Duan, Zheyuan Liu, Xueyang Li, Meng Jiang, Michael Lemmon, Wei Jin, and Yiyu Shi. Incorporating Rather Than Eliminating: Achieving Fairness for Skin Disease Diagnosis Through Group-Specific Experts. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 284–294, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-05185-1. doi: 10.1007/978-3-032-05185-1_28.

## Appendix A. Additional Methods

**ADNI Dataset.** Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

**Cohort Demographics.** In our experiments, we used random sampling without explicitly stratifying based on patient demographics. We have analyzed the demographic distributions of the training and test sets and found no statistical significant difference in any demographic dimension (Table 2).

Table 2: Comparison of demographic characteristics between Training set and Test set.

| Characteristic | Train Set (n=541) | Test Set (n=100) | p-value |
|---|---|---|---|
| **Age** | 71.5 ± 7.5 | 70.5 ± 7.1 | 0.23 |
| **Years of Education** | 16.1 ± 2.7 | 16.6 ± 2.4 | 0.12 |
| **Gender** | | | 0.63 |
| - Male | 247 (54.8%) | 58 (58.0%) | |
| - Female | 204 (45.2%) | 42 (42.0%) | |
| **Race** | | | 0.19 |
| - White | 423 (93.8%) | 96 (96.0%) | |
| - Black | 12 (2.7%) | 1 (1.0%) | |
| - Asian | 6 (1.3%) | 0 (0.0%) | |
| - Other | 10 (2.2%) | 3 (1.0%) | |
| **APOE4 status** | | | 0.58 |
| - 0 | 241 (53.4%) | 55 (55.0%) | |
| - 1 | 164 (36.4%) | 32 (32.0%) | |
| - 2 | 46 (10.2%) | 13 (13.0%) | |
| **APOE2 status** | | | 0.45 |
| - 0 | 416 (92.2%) | 95 (95.0%) | |
| - 1 | 35 (7.8%) | 5 (5.0%) | |

**Model Implementation Details.** As mentioned in Section 2.2, we report the final hyperparameters of the models below:

**Basic Model**

```
sklearn.linear_model.LogisticRegression
(C=0.1, penalty='l2', solver='liblinear')
```

**Advanced Model**

```
sklearn.linear_model.LogisticRegression
(C=0.01, penalty='l2', solver='liblinear')
```

**Triage Model**

```
sklearn.svm.SVC
(C=10, kernel='rbf', gamma='auto')
```

## Appendix B. Additional Results

**ROC and PR Curves.** We further compare performance of our proposed approach against baselines by plotting ROC and PR curves on the test set. As shown in Figure 5, the ROC curve of the Triage Model consistently arches higher and closer to the optimal top-left corner compared to all baselines (though it does not dominate all baselines). Similarly, the model's PR curve also maintains a higher elevation across different recall values (Figure 6), demonstrating its ability to sustain generally higher precision than the baselines and single model. Notably, the PR curve reveals that our triage system maintains perfect precision up to a recall of nearly 60%. This is clinically significant as it indicates that the top half of patients identified by the model as high-risk are classified with extremely high confidence, minimizing the risk of false positives in this critical subgroup.

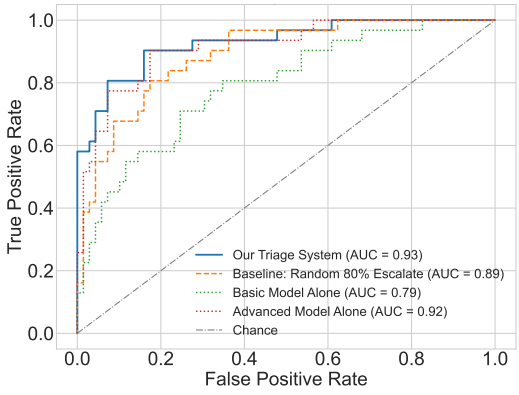


Figure 5: ROC curves for all models on the test set.

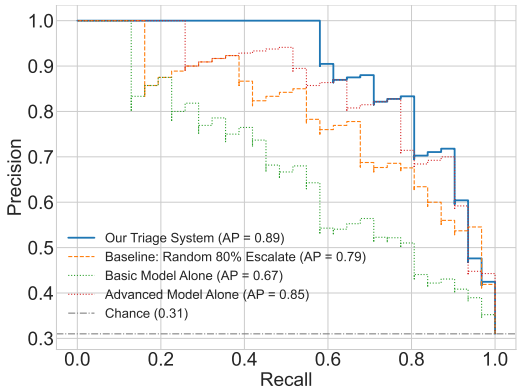


Figure 6: PR curves for all models on the test set.

**Statistical Significance of Performance Differences.** We conducted bootstrapped resampling tests to compare the performance of our approach with the Basic Model alone, the Advanced Model alone, and several baselines, with AUROC differences detailed in Table 3. The Triage Model demonstrates a significantly higher AUROC than both the Basic Model alone ($\Delta$AUROC = +0.1370, p=0.0010) and the baseline of randomly escalating 80% of patients ($\Delta$AUROC = +0.0388, p=0.0330). Although the Triage Model also numerically outperforms the more resource-intensive Advanced Model and the remaining baselines, these smaller performance gains were not statistically significant (p>0.05).

Table 3: Statistical comparison of AUROC scores of the triage framework against baselines.

| Comparison Model | $\Delta$AUROC | p-value | Significant? |
|---|---|---|---|
| Basic Model Alone | +0.1370 | 0.0010 | Yes |
| Advanced Model Alone | +0.0131 | 0.1010 | No |
| Baseline: Random 80 Escalate | +0.0388 | 0.0330 | Yes |
| Baseline: Escalate 80 Highest Prob | +0.0112 | 0.0890 | No |
| Baseline: Escalate 80 Most Uncertain | +0.0028 | 0.5020 | No |

**Comparative Cost-Effectiveness Analysis.** We further compared various baselines in the cost-AUROC trade-off analysis (Figure 7). The triage system's performance trajectory (blue) remains consistently superior to Random Escalation baselines, but remained close to the other baselines, and it notably overlaps with the baseline that escalates based on the most uncertain cases.

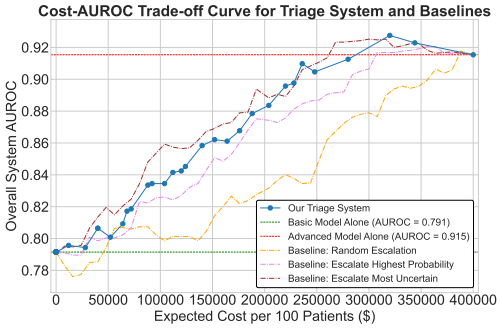


Figure 7: Cost-AUROC trade-off curves comparing the triage system with baselines.

**Fairness Analysis.** To assess the fairness of our Triage Model, we analyzed the demographic distribution of the patient groups recommended for escalation versus those not escalated (at the $\tau = 0.05$ threshold). We found no statistically significant difference in gender, race, or education level between the groups (Table 4). The one factor with a significant difference was APOE4 status (p=0.02), which is a primary clinical risk factor for AD. This suggests the model's triage decisions are driven by known clinical factors, not demographic bias. In future work, we will expand our study to a larger cohort and systematically evaluate and mitigate potential fairness issues (Xu et al., 2026).

Table 4: Comparison of demographic characteristics between escalated and non-escalated groups.

| Characteristic | Escalated Group (%) | Non-Escalated Group (%) | p-value |
|---|---|---|---|
| **Education Level** | | | 0.93 |
| - 16+ years | 40% | 42% | |
| - 12-15 years | 35% | 38% | |
| - < 12 years | 25% | 20% | |
| **Gender** | | | 0.77 |
| - Male | 55% | 58% | |
| - Female | 45% | 42% | |
| **Race** | | | 0.97 |
| - White | 85% | 88% | |
| - Black | 10% | 8% | |
| - Asian | 3% | 2% | |
| - Other | 2% | 2% | |
| **APOE4 Status** | | | 0.02 |
| - 0 | 30% | 55% | |
| - 1 | 45% | 30% | |
| - 2 | 25% | 15% | |
| **APOE2 Status** | | | 0.81 |
| - 0 | 10% | 12% | |
| - 1 | 90% | 88% | |