
Validating Deep Models for Alzheimer’s ^{18}F -FDG PET Diagnosis Across Populations: A Study with Latin American Data

Hugo Massaroli^{1,2} Hernan Chaves³ Pilar Anania³ Mauricio Farez³ Emmanuel Iarussi¹ Viviana Siless¹

Abstract

Deep learning models have shown strong performance in diagnosing Alzheimer’s disease (AD) using neuroimaging data, particularly ^{18}F -FDG PET scans, with training datasets largely composed of North American cohorts such as those in the Alzheimer’s Disease Neuroimaging Initiative (ADNI). However, their generalization to underrepresented populations remains under-explored. In this study, we benchmark convolutional and Transformer-based models on the ADNI dataset and assess their generalization performance on a novel Latin American clinical cohort from the FLENI Institute in Buenos Aires, Argentina. We show that while all models achieve high AUCs on ADNI (up to .96, .97), their performance drops substantially on FLENI (down to .82, .80, respectively), revealing a significant domain shift. The tested architectures demonstrated similar performance, calling into question the supposed advantages of transformers for this specific task. Through ablation studies, we identify per-image normalization and a correct sampling selection as key factors for generalization. Occlusion sensitivity analysis further reveals that models trained on ADNI, generally attend to canonical hypometabolic regions for the AD class, but focus becomes unclear for the other classes and for FLENI scans. These findings highlight the need for population-aware validation of diagnostic AI models and motivate future work on domain adaptation and cohort diversification.

1. Introduction

Alzheimer’s disease (AD) is the most common cause of dementia, accounting for an estimated 60–70% of the 57 million dementia cases worldwide as of 2021 (World Health Organization, 2021). With aging populations, this number is projected to rise sharply in the coming decades. Early and accurate diagnosis is essential for managing disease progression and informing treatment strategies. ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) has proven to be a reliable biomarker of cerebral glucose metabolism, revealing characteristic patterns of hypometabolism associated with AD (Chételat et al., 2020).

Recent years have witnessed the application of deep learning methods to neuroimaging for automated AD diagnosis (Ding et al., 2019). Models trained on large datasets such as the Alzheimer’s Disease Neuroimaging Initiative (ADNI) have achieved high classification accuracy. However, there is growing concern that such models may not generalize well to demographically or clinically distinct populations (Arora et al., 2023; Yang et al., 2024).

In this work, we explore this challenge by comparing the performance of three deep learning models—a convolutional neural network (CNN), a transformer-based architecture, and a lightweight ResNet (He et al., 2016) variant—trained on ADNI and evaluated on a novel Latin American dataset from the Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia (FLENI) Institute in Argentina. This cross-cohort evaluation enables a realistic assessment of model robustness in diagnostic AI applications.

Our contributions are threefold:

- We benchmark and compare three architectures under matched training settings, reporting both in-distribution and out-of-distribution performance.
- We conduct ablation studies on preprocessing strategies and input representations to identify key factors impacting generalization.
- We provide visual diagnostic analyses to interpret model decisions across domains, revealing differences in spatial attention patterns.

¹Business School, Di Tella University, Argentina

²Departamento de Computación, Universidad de Buenos Aires, Buenos Aires, Argentina ³Fleni Institute, Buenos Aires, Argentina. Correspondence to: Viviana Siless <viviana.siless@utdt.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Our findings expose consistent performance degradation on the FLENI dataset and support the claim that standard benchmarks alone are insufficient for evaluating clinically deployable models.

2. Related Work

Research on automated Alzheimer’s disease diagnosis using neuroimaging has been active for over a few decades. Early approaches focused on handcrafted features derived from structural MRI and FDG-PET, including hippocampal atrophy and hypometabolic patterns in temporoparietal regions—both established biomarkers of AD progression (Frisoni et al., 2010; Weiner et al., 2015).

More recently, deep learning methods have achieved state-of-the-art performance by automatically learning representations from raw imaging data. In particular, Convolutional Neural Networks (CNNs) have been the dominant architecture in AD neuroimaging tasks, particularly for image classification and disease stage prediction (Ding et al., 2019). These models are well-suited for capturing local spatial patterns and have demonstrated robustness in both binary and multi-class AD classification problems.

More recently, Transformer-based architectures (Vaswani et al., 2017), originally developed for natural language processing, have recently gained traction in medical imaging due to their ability to model long-range dependencies via self-attention mechanisms (Shamshad et al., 2023; He et al., 2023; Li et al., 2023). Unlike CNNs, which focus on localized receptive fields, transformers can capture global contextual relationships, which is particularly valuable in neuroimaging where subtle and spatially distributed patterns may be predictive of disease. However, to the best of our knowledge, Transformer-based models have not yet been introduced into FDG-PET analysis for Alzheimer’s disease.

Whether models trained on one population can generalize effectively to diverse clinical cohorts remains an open and critical question in the field (Arora et al., 2023; Yang et al., 2024). Most AD classification studies are conducted on homogeneous datasets such as ADNI, which does not reflect the diversity of real-world populations. Evaluating models on independent cohorts, particularly from underrepresented regions like Latin America, is essential for assessing their true clinical utility.

Despite growing interest in fair and generalizable AI systems, to our knowledge, most studies in AD diagnosis continue to rely on demographically homogeneous datasets. This poses significant limitations in regions where population health characteristics, imaging protocols, and health-care infrastructure differ from those represented in benchmark datasets. To our knowledge, this is the first study to explicitly evaluate the generalization of both CNN and

transformer-based architectures trained on ADNI to a Latin American cohort using FDG-PET data.

3. Data

We used multiple FDG-PET datasets to assess model performance and generalization across heterogeneous cohorts: ADNI, FLENI100, and FLENI600.

3.1. ADNI Dataset

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) is a longitudinal, multicenter study aimed at developing clinical, imaging, genetic, and biochemical biomarkers for the early detection of Alzheimer’s disease (AD). We included subjects from ADNI1, ADNI-GO, ADNI2, and ADNI3, focusing on FDG-PET scans.

The dataset comprises 3,762 pre-processed image records from 1,687 unique subjects. Each subject contributed multiple scans due to longitudinal follow-up. The image volumes are composed of 96 DICOM slices each, representing 3D brain scans of dimensions $160 \times 160 \times 96$ voxels with an isotropic resolution of 1.5 mm.

ADNI diagnostic labels are assigned clinically at each visit and span five classes: cognitively normal (CN), subjective memory concerns (SMC), early mild cognitive impairment (EMCI), mild cognitive impairment (MCI), late mild cognitive impairment (LMCI), and Alzheimer’s dementia (AD). For this study, we excluded SMC and merged EMCI, LMCI, and MCI into a unified MCI category. To facilitate comparison with FLENI’s biomarker-driven labeling, we created two label variants:

- `visit953`: Labels are based on the clinical diagnosis closest to the FDG-PET scan, with two classes:
 - **AD**: Subjects diagnosed with Alzheimer’s disease (AD) at scan time, or with mild cognitive impairment (MCI) who later converted to AD.
 - **Non-AD**: Cognitively normal (CN) subjects and MCI cases that did not convert to AD, grouped to match FLENI’s non-AD biomarker profile.

Each subject contributes one scan. For MCI-to-AD converters, we used the latest scan labeled AD; for CN-to-MCI converters, the earliest scan labeled CN. MCI subjects who neither converted nor originated from CN were excluded.

- `last`: Labels reflect the diagnosis at the subject’s final recorded visit, with three classes: CN, MCI, and AD.

Table 1. Demographics and class distribution across datasets.

Dataset	Labeling	Gender	AD	MCI	CN	Non-AD	Total
ADNI	last	Male	299	428	204	—	931
		Female	206	306	231	—	743
ADNI	visit953	Male	291	—	—	229	520
		Female	195	—	—	238	433
FLENI	FLENI100	Male	33	—	—	24	57
		Female	27	—	—	16	43
FLENI	FLENI600	Male	147	—	—	154	301
		Female	173	—	—	120	293

3.2. FLENI Datasets

The FLENI datasets originate from the Fundación para la Lucha contra las Enfermedades Neurológicas de la Infancia (FLENI) in Argentina. These datasets include patients evaluated with both PIB and FDG PET tracers. Image acquisition was performed using a General Electric Discovery 690 PET/CT scanner.

Each subject underwent a two-step scanning protocol. First, 13.44 mCi of ^{11}C -PIB was administered intravenously, followed by a 50-minute uptake period and a 20-minute scan. Thirty minutes later, 6.49 mCi of ^{18}F FDG was administered, followed by a 30-minute rest and a subsequent 20-minute scan. Each FDG study consists of 47 DICOM slices forming a 3D image of size $128 \times 128 \times 47$ with voxel dimensions of $2 \times 2 \times 3.27$ mm (non-isotropic). When using the Transformer model with input samples containing 77 axial slices, the data is resized to $128 \times 128 \times 77$ voxels using nearest-neighbor interpolation.

FLENI100: This subset consists of 100 studies with biomarker-based diagnoses. AD status is determined by positive cerebrospinal fluid (CSF) beta-amyloid and phosphorylated tau levels or corresponding PIB PET positivity, enabling early detection before dementia symptoms appear. Control subjects were biomarker-negative. Both AD and Non-AD cohorts may include MCI patients.

FLENI600: This larger dataset includes 594 studies labeled based on PET imaging alone. AD subjects exhibit characteristic cortical FDG hypometabolism and PIB retention. Controls present with normal FDG uptake and no PIB signal. Age information is not available for FLENI600. As in FLENI100, both cohorts may include MCI subjects.

3.3. Cohort Demographics and Label Distribution

Table 1 summarizes the gender and diagnostic breakdown for each dataset and label strategy. Age statistics for ADNI and FLENI100 are available, with ADNI subjects ranging from 55 to 96 years and FLENI100 from 43 to 90. FLENI600 age data is not recorded.

4. Methods

Preprocessing included skull stripping and resampling to a uniform voxel grid of $128 \times 128 \times 77$, corresponding to the sagittal, coronal, and axial planes, respectively. Intensity values were normalized using z-score normalization either globally or per subject. For CNN-based models, we extracted 16 equidistant axial slices. For Transformers, we used full 3D volumes with orthogonal plane projections.

4.1. Model Architectures

We implemented and evaluated three neural network architectures:

- **Inception CNN:** A 2D convolutional model based on InceptionV3 operating on 16 axial FDG-PET slices arranged into a 4×4 grid, as in (Ding et al., 2019). It uses ReLU activations, dropout, and softmax output.
- **Transformer:** Adapted from Medical Transformer (Jun et al., 2024), it integrates a ResNet-18 encoder and multi-head self-attention across anatomical planes with positional encodings. Each brain slice is represented by a single token with 64 features. Tokens from all three anatomical planes are concatenated, resulting in a sequence of 333 tokens per image. The transformer layer employs 4 attention heads to contextualize the token sequence. Following this, the architecture includes a dropout layer, two fully connected layers, and a softmax output. To ensure a fair comparison with the other models, this architecture can also be configured to use 16 axial slices instead of 77.
- **Pruned ResNet (P-ResNet):** A streamlined model derived from the Transformer, using 16 axial slices and omitting the transformer encoder and sagittal/coronal tokens. Each slice is embedded to a single scalar (reduced from 64), resulting in a 16-dimensional input to the final fully connected layer. This architecture has 718,000 parameters, making it significantly lighter than the Inception and Transformer models.

All models were implemented in PyTorch 2.0.1 and trained using the Adam optimizer. Data augmentation included rotation, flipping, intensity jittering, and Gaussian noise.

4.2. Training and Validation Protocol

Models were trained on ADNI using 10-fold cross-validation with stratified splits. The best model was selected based on validation AUC-ROC. For generalization assessment, models were evaluated on the FLENI datasets without fine-tuning.

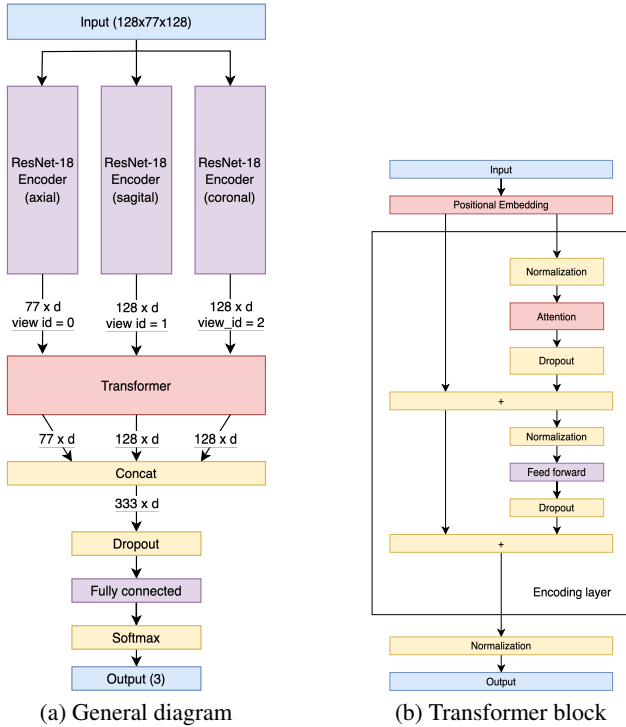


Figure 1. Schematic of the Transformer architecture. Multi-view ResNet encoders process axial, sagittal, and coronal slices. Features are flattened into patches with positional encoding and passed to a transformer encoder.

Hyperparameters were selected via grid search. The Inception model used a learning rate of 0.0001 and a batch size of 32. Transformers and P-ResNet used a learning rate of 0.0005 and a batch size of 8. Cross-entropy loss was weighted by inverse class frequency. Early stopping was applied with a patience of 30 epochs. This is summarized in Table 2.

Experiments were conducted on a computing server equipped with two NVIDIA A100 80GB GPUs and 256 GB of RAM.

Table 2. Model hyperparameters and configurations.

Model	LR	Batch Size	Dropout	Epochs	Patience
Inception	0.0001	32	0.6	100	30
Transformer	0.0005	8	0.4	100	30
P-ResNet	0.0005	8	0.4	100	30

4.3. Evaluation Metrics

We report the area under the ROC curve (AUC), accuracy, sensitivity, and specificity for both ADNI (in-distribution) and FLENI (out-of-distribution) evaluations. Occlusion sensitivity maps were used for interpretability.

5. Results

5.1. Model Performance

We report all experimental configurations for Inception, Transformer and P-ResNet models in Table 3, including normalization strategies, slice count, and AUC values across datasets.

In the context of sample selection strategies, **first** refers to using only the first scan per subject in the validation and test sets, ensuring that no individual contributes multiple data points during evaluation. **First w/train** extends this constraint to the training set as well, using only the first scan per subject throughout all data partitions. This strategy aims to evaluate model generalization in a more realistic clinical scenario where repeated measures are not available.

This comprehensive table merges all configurations and helps illustrate the trade-offs across normalization, model type, input strategy, and labeling. Best AUCs on FLENI datasets are highlighted in bold. A summary of top-performing settings remains in Table 3.

5.2. Stability and Convergence

Training behavior varied across architectures. The Transformer model, when using 16 slices, exhibited less variation across folds and showed signs of overfitting, converging in an average of 19 epochs. It was followed by the Inception model, with an average best epoch of 24. In contrast, the Transformer (77 slices) and ResNet architectures showed fewer signs of overfitting, with an average convergence epoch of 30 and 46 epochs each.

5.3. Ablation Studies

To isolate the impact of preprocessing and input configuration, we conducted a series of ablation experiments on every model. Table 3 summarizes how modifications such as slice selection strategy, image selection strategy, and normalization approach influenced generalization.

Table 3. Merged AUC ROC Results for Inception, Transformer and P-ResNet Models. Sample selection strategies: **first** uses the first scan per subject for validation and testing; **first w/train** applies the same restriction to training as well.

Model	Labeling	Classes	Slices	Normalization	ADNI Test	FLENI100	FLENI600
Inception	last	3	16	z-score global (first w/train)	0.83 (0.02)	0.63 (0.03)	0.72 (0.03)
	last	3	16	z-score (first)	0.82 (0.02)	0.69 (0.09)	0.78 (0.04)
	last	3	16	z-score (first w/train)	0.83 (0.02)	0.71 (0.05)	0.81 (0.03)
	last	3	16	min-max (first w/train)	0.82 (0.02)	0.70 (0.03)	0.80 (0.02)
	visit953	2	16	z-score global	0.95 (0.01)	0.64 (0.04)	0.68 (0.03)
	visit953	2	16	z-score per-image	0.96 (0.01)	0.73 (0.04)	0.82 (0.02)
	visit953	2	16	min-max	0.96 (0.01)	0.73 (0.03)	0.81 (0.02)
Transformer	last	3	77	min-max	0.84 (0.02)	0.70 (0.03)	0.78 (0.03)
	last	3	16	min-max	0.82 (0.01)	0.68 (0.04)	0.78 (0.01)
	visit953	2	77	min-max	0.95 (0.01)	0.76 (0.02)	0.81 (0.01)
	visit953	2	16	min-max	0.96 (0.01)	0.72 (0.03)	0.81 (0.02)
	last	3	77	z-score per-image	0.84 (0.02)	0.71 (0.03)	0.79 (0.02)
	last	3	16	z-score per-image	0.81 (0.01)	0.66 (0.04)	0.78 (0.02)
	visit953	2	77	z-score per-image	0.95 (0.01)	0.74 (0.02)	0.81 (0.02)
	visit953	2	16	z-score per-image	0.95 (0.01)	0.71 (0.02)	0.79 (0.02)
P-ResNet	last	3	16	min-max	0.84 (0.01)	0.74 (0.02)	0.81 (0.01)
	visit953	2	16	min-max	0.96 (0.01)	0.72 (0.03)	0.80 (0.01)
	last	3	16	z-score per-image	0.84 (0.01)	0.74 (0.03)	0.82 (0.01)
	visit953	2	16	z-score per-image	0.97 (0.00)	0.71 (0.02)	0.80 (0.01)

The results emphasize the importance of both spatial context and normalization scheme. Full axial coverage and per-image z-score normalization yielded the most robust cross-domain performance. In the Transformer model, reducing spatial input weakened generalization, especially on the FLENI dataset.

5.4. Visual Diagnostics

To further understand model behavior and domain-specific biases, we employed occlusion sensitivity analysis on the Inception model. This method involves masking out spatial patches of the input and quantifying the change in prediction confidence, thereby revealing regions critical for classification.

Figure 2 displays occlusion maps from representative samples. In ADNI, we highlight three diagnostic classes (CN, MCI, AD), whereas FLENI 600 reflects binary classification (Non-AD vs. AD).

In ADNI, occlusion-based relevance scores tend to be stronger and more concentrated, suggesting more confident and localized decision patterns. Surprisingly, the regions that the network attends to in FLENI differ from those in ADNI, with more diffuse or inconsistent activation. This indicates possible structural or intensity mismatches between cohorts that may challenge the model’s learned representations. These qualitative findings mirror the observed

quantitative performance drop and reinforce the clinical importance of validating models across diverse populations.

6. Discussion

Our experiments show that while deep learning models for Alzheimer’s disease classification can achieve strong performance within a controlled training domain (e.g., ADNI), their generalization to external, demographically distinct cohorts remains a major challenge. All three models—Inception CNN, Transformer, and pruned ResNet—showed reduced performance on the FLENI dataset, with AUCs dropping by over 10 percentage points in some cases. This domain shift underscores the need to assess model robustness not just on held-out data from the same source, but across clinically realistic, geographically diverse populations.

All evaluated architectures—Inception, Transformer, and ResNet—achieved strong performance on both the in-distribution (ADNI) dataset and the out-of-distribution (FLENI) dataset. The best Transformer configuration reached an AUC of 0.95 on ADNI and 0.76/0.81 on FLENI100/FLENI600. P-ResNet followed closely with 0.84/0.74/0.82, showing strong cross-cohort generalization despite its compact architecture. Inception achieved comparable performance, notably with a best result of 0.96 on ADNI and 0.73/0.82 on FLENI. While no single model

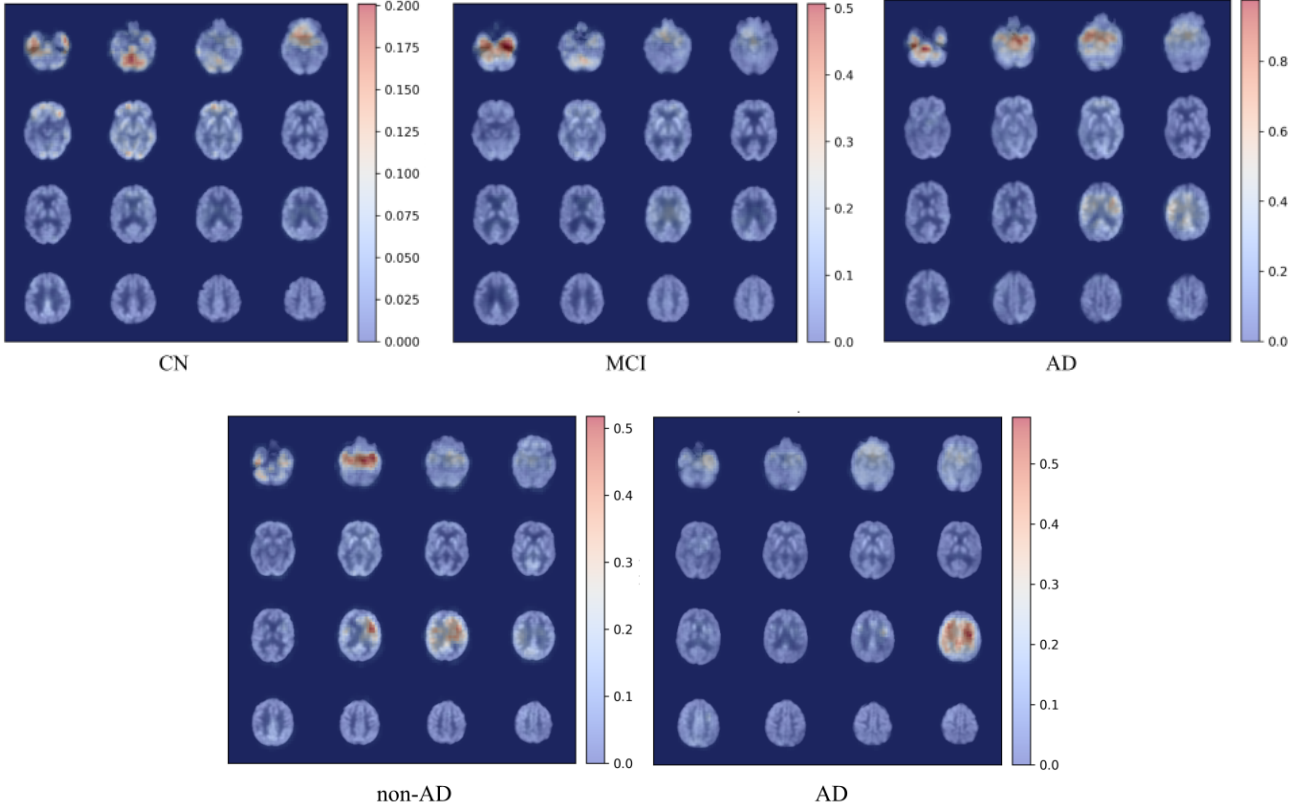


Figure 2. Occlusion sensitivity maps from the Inception model. Top: ADNI samples (CN, MCI, AD). Bottom: FLENI 600 samples (non-AD, AD). The highlighted regions indicate areas with the greatest impact on model predictions.

outperformed all others in every setting, Inception offered the strongest performance on ADNI without compromising generalization. These results highlight that effective generalization across cohorts does not necessarily require the most complex model, and underscore the importance of systematic cross-dataset validation.

Ablation studies further revealed that full-slice input and per-image normalization are key for generalization. Simplifying spatial input or altering preprocessing steps (e.g., global z-score normalization) consistently degraded FLENI performance. These insights emphasize the importance of preprocessing consistency and spatial completeness in training models intended for clinical generalization.

Saliency analyses provided additional confirmation of these findings, showing stronger attention to regions in ADNI, but attenuated or shifted attention in FLENI. These differences likely reflect both demographic variability and acquisition protocol discrepancies between cohorts.

7. Conclusion

This work presents a comprehensive evaluation of deep learning models for Alzheimer’s disease diagnosis using

FDG-PET scans, with a focus on validating generalization across geographic and demographic boundaries. Our results show that all evaluated architectures—Inception, Transformer, and P-ResNet—performed comparably, achieving high AUC scores on ADNI.

However, when tested on the out-of-distribution FLENI datasets, every model exhibited at least a 10-point drop in AUC, despite using optimized configurations. This consistent drop highlights the risks of relying solely on in-domain validation and underscores the necessity of cross-cohort benchmarking when deploying diagnostic models in real-world clinical environments.

Future work should investigate training with more diverse global datasets, as well as domain adaptation and harmonization techniques to reduce performance degradation and improve fairness across underrepresented populations.

References

- Arora, A., Alderman, J. E., Palmer, J., Ganapathi, S., Laws, E., McCradden, M. D., Oakden-Rayner, L., Pfohl, S. R., Ghassemi, M., McKay, F., Treanor, D., Ros-tamzadeh, N., Mateen, B., Gath, J., Adebajo, A. O.,

- Kuku, S., Matin, R., Heller, K., Sapey, E., Sebire, N. J., Cole-Lewis, H., Calvert, M., Denniston, A., and Liu, X. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929–2938, October 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02608-w. URL <http://dx.doi.org/10.1038/s41591-023-02608-w>.
- Chételat, G., Arbizu, J., Barthel, H., Garibotto, V., Law, I., Morbelli, S., van de Giessen, E., Agosta, F., Barkhof, F., Brooks, D. J., et al. Amyloid-pet and 18f-fdg-pet in the diagnostic investigation of alzheimer’s disease and other dementias. *The Lancet Neurology*, 19(11):951–962, 2020.
- Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., Lituiev, D., Copeland, T. P., Aboian, M. S., Mari Aparici, C., et al. A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain. *Radiology*, 290(2):456–464, 2019.
- Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M. The clinical use of structural mri in alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, February 2010. ISSN 1759-4766. doi: 10.1038/nrneurol.2009.215. URL <http://dx.doi.org/10.1038/nrneurol.2009.215>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., and Shen, D. Transformers in medical image analysis. *Intelligent Medicine*, 3(1):59–78, 2023.
- Jun, E., Jeong, S., Heo, D.-W., and Suk, H.-I. Medical transformer: Universal encoder for 3-d brain mri analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17779–17789, December 2024. ISSN 2162-2388. doi: 10.1109/tnnls.2023.3308712. URL <http://dx.doi.org/10.1109/TNNLS.2023.3308712>.
- Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., and Zhou, S. K. Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, 85:102762, 2023.
- Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. Transformers in medical imaging: A survey. *Medical image analysis*, 88:102802, 2023.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. Impact of the alzheimer’s disease neuroimaging initiative, 2004 to 2014. *Alzheimer’s & Dementia*, 11(7):865–884, 2015.
- World Health Organization. Dementia, 2021. URL <https://www.who.int/news-room/fact-sheets/detail/dementia>.
- Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., and Ghassemi, M. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10):2838–2848, June 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03113-4. URL <http://dx.doi.org/10.1038/s41591-024-03113-4>.