SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 3, ISSUE 12, 2025

MINT-NET: TOWARD INTELLIGIBLE NEURO-AI FOR MULTIMODAL BRAIN DISORDER CLASSIFICATION VIA CROSS-MODAL ATTENTION AND NEURAL REPRESENTATION LEARNING

Maaz Allah[*1], Altamash Afridi[1], Muhammad Imran Ali[3], Habibullah[4]

[*1,2]*Department of Computer Systems Engineering, University of Engineering and Technology Peshawar*
[3]*Department of Informatics Engineering University of Coimbra, Portugal*
[4]*Department of Mechanical Engineering University of Engineering and Technology Peshawar*

[*1]maaz.uthman@gmail.com, [2]ta.afridii@gmail.com, [3]mali@dei.uc.pt , [4]habibrahi3543@gmail.com

DOI: http://doi.org/10.5281/zenodo.20507929

Keywords
Alzheimer's disease, multimodal deep learning, cross-modal attention, interpretable AI, neuroimaging, representation learning, ADNI

Article History
Received: 03 October 2025
Accepted: 15 November 2025
Published: 30 December 2025

Corresponding Author: *
Maaz Allah

Abstract

Classifying Accurately distinguishing Alzheimer's Disease (AD) from normal cognitive aging remains a major challenge in neuroimaging research. Although deep learning models have achieved remarkable classification performance, their black-box nature continues to limit clinical trust and scientific interpretability. In high-stakes medical settings, predictive systems must not only provide accurate diagnoses, but also offer meaningful insight into the reasoning behind their decisions. To address this challenge, we propose MINT-Net (Multimodal Intelligible Neural Network), a framework designed to jointly optimize predictive accuracy and neurobiological interpretability. MINT-Net integrates structural MRI, FDG-PET, and clinical information through a cross-modal multi-head attention mechanism that models complex interactions across modalities rather than relying on simple feature concatenation. In addition, the framework learns a shared latent representation space using supervised contrastive learning and center loss, encouraging representations that are both discriminative and clinically meaningful [10], [11]. To improve interpretability, we introduce a hierarchical intelligibility module that combines gradient-based attention visualization with SHAP-based clinical feature attribution [12], [13]. This allows interpretation at multiple levels, ranging from individual predictions to population-level biomarker analysis. We evaluated MINT-Net on 515 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset using rigorous 5-fold cross-validation [31]. The proposed framework achieved 96.12% classification accuracy, 96.92% sensitivity, and an AUC-ROC of 0.984, outperforming six state-of-the-art baseline methods. Ablation analysis further demonstrated that cross-modal attention improved performance by 2.67% compared to direct feature concatenation, while supervised contrastive learning improved latent space separability by approximately 35%. Attention visualizations consistently highlighted the hippocampus, posterior cingulate cortex, and temporoparietal junction, regions strongly associated with AD pathology [1], [2]. Overall, MINT-Net provides not only accurate diagnostic predictions, but also interpretable neurobiological insight into disease-related representations. The framework represents a step toward bridging predictive Neuro-AI with scientifically grounded understanding of brain disorders neuroimaging.

INTRODUCTION

Alzheimer's Disease (AD) is the most prevalent neurodegenerative disorder worldwide, currently affecting more than 55 million individuals, with prevalence expected to nearly triple by 2050 [1]. Early and reliable diagnosis is critical for timely therapeutic intervention; however, current clinical diagnosis still depends heavily on cognitive assessment and invasive biomarker procedures. Neuroimaging modalities, particularly structural Magnetic Resonance Imaging (MRI) and fluorodeoxyglucose Positron Emission Tomography (FDG-PET), provide non-invasive insight into structural degeneration and metabolic dysfunction, making them central to modern AD research frameworks [2].

Deep learning has emerged as a powerful approach for automated AD classification from neuroimaging data [3]. Convolutional Neural Networks (CNNs) applied to structural MRI have demonstrated classification accuracies exceeding 90% [4], [5]. More recently, multimodal frameworks integrating MRI, PET, and clinical variables have achieved even stronger predictive performance, with some approaches approaching near-perfect accuracy [6].

Despite these advances, a critical limitation remains unresolved. Most high-performing models function as black boxes, providing limited explanation for how predictions are generated. This lack of interpretability creates several challenges. First, clinicians are often reluctant to rely on systems that cannot justify their decisions using biologically meaningful reasoning [7]. Second, black-box models provide limited scientific insight into disease mechanisms and progression [8]. Third, the absence of transparent reasoning makes it difficult to determine whether models are learning genuine disease patterns or merely exploiting hidden biases within datasets [9].

Contributions

To address these limitations, we propose MINT-Net, a framework explicitly designed to bridge predictive modeling and intelligible neural representation learning.

Our main contributions are summarized as follows:

We introduce a cross-modal multi-head attention fusion mechanism that integrates structural MRI, FDG-PET, and clinical data by explicitly modeling interactions across modalities rather than relying on direct feature concatenation [35].

We employ supervised contrastive learning with center loss to encourage strong separation between AD and cognitively normal (CN) representations within the latent space, producing features that are both discriminative and neurobiologically meaningful [10], [11].

We develop a hierarchical intelligibility module that combines Grad-CAM-based spatial attention visualization with SHAP-based clinical feature attribution, enabling interpretation at both prediction and population levels [12], [13].

We conduct extensive evaluation on 515 ADNI subjects using rigorous 5-fold cross-validation [31], demonstrating strong predictive performance while showing that learned attention patterns align with established AD neurobiology [1], [2].

II. Related Work

A. Deep Learning for AD Classification

Early deep learning approaches for Alzheimer's Disease classification primarily focused on unimodal MRI analysis. Suk et al. [5] introduced hierarchical feature learning using autoencoder-based multimodal fusion, while Liu et al. [4] demonstrated the effectiveness of deep residual networks for modeling structural atrophy progression across the AD spectrum. Spasov et al. [14] later proposed a parameter-efficient 3D CNN architecture capable of achieving competitive performance with lower computational complexity. Wen et al. [15] further contributed reproducible evaluation protocols that helped standardize comparison across CNN-based AD classification methods.

Research on brain age prediction further expanded the role of deep learning in neuroimaging. Cole et al. [16] demonstrated that CNN-based brain age models could identify accelerated aging patterns associated with neurodegenerative disease. Similarly, Klöppel et al. [17] established a strong relationship between structural MRI

abnormalities and AD diagnosis using pattern recognition techniques.

B. Multimodal Fusion Approaches

As unimodal approaches began reaching performance limitations, attention shifted toward multimodal fusion strategies. Mathur et al. [18] employed cross-modal attention to integrate MRI, genetic, and clinical data, achieving classification accuracy close to 97%. Ahmad et al. [6] proposed NeuroNet-AD, a framework combining CBAM-attention ResNet architectures with meta-guided cross-attention mechanisms, reporting performance exceeding 99% on binary classification tasks. More recently, Vision Transformer-based architectures have also been adapted for neuroimaging applications, demonstrating strong performance when sufficient training data is available [21], [22].

Despite these advances, most multimodal methods primarily optimize predictive accuracy rather than interpretability. In many cases, explanatory mechanisms remain post-hoc and may not accurately reflect the model's true reasoning process [7], [8].

C. Interpretability in Neuro-AI

Interpretability techniques such as Grad-CAM [12], SHAP [13], and LIME [24] are widely used in medical AI to explain deep learning predictions. However, these approaches are typically applied after model training and may not always provide reliable or biologically meaningful explanations [7], [8]. Earlier visualization methods such as deconvolutional networks [25] and Layer-wise Relevance Propagation (LRP) [23] attempted to improve transparency by highlighting influential input regions. Grad-CAM later became one of the most widely adopted approaches for visualizing class-discriminative brain regions [12], while SHAP introduced a game-theoretic framework for estimating feature importance [13].

More integrated interpretability strategies have recently emerged. Gazzarrini et al. [8] proposed concept bottleneck models that constrain predictions through interpretable intermediate concepts. Similarly, attention-based Neuro-AI frameworks have demonstrated stronger neurobiological plausibility in their explanations [27]. These studies collectively emphasize the growing need for models that unify statistical learning with structured neurobiological reasoning [26].

D. Representation Learning in Medical Imaging

Representation learning has become increasingly important for understanding latent disease structure in medical imaging. Supervised contrastive learning, introduced by Khosla et al. [10], extended self-supervised contrastive methods into supervised settings, significantly improving representation quality. Wen et al. [11] demonstrated that center loss could further improve discriminative feature learning by encouraging compact class distributions. Recent studies have shown that appropriately trained deep learning models can reveal previously unknown neuropathological patterns [28]. In parallel, multimodal fusion architectures [29] and brain-inspired AI frameworks [30] continue to advance the integration of computational neuroscience and representation learning in medical imaging.

III. Methodology

A. Overview

MINT-Net combines four tightly connected components working together within a unified framework, illustrated in Fig. 1. First, modality-specific encoders extract meaningful representations from structural MRI, FDG-PET, and clinical data. Next, a cross-modal multi-head attention fusion module models the interactions between these heterogeneous modalities instead of treating them independently. The learned features are then projected into a shared latent representation space through supervised contrastive learning and center loss, encouraging both discriminative and clinically meaningful embeddings. Finally, a hierarchical intelligibility module provides interpretation at multiple levels, allowing the model's reasoning process to be examined from spatial brain regions to clinical feature importance.

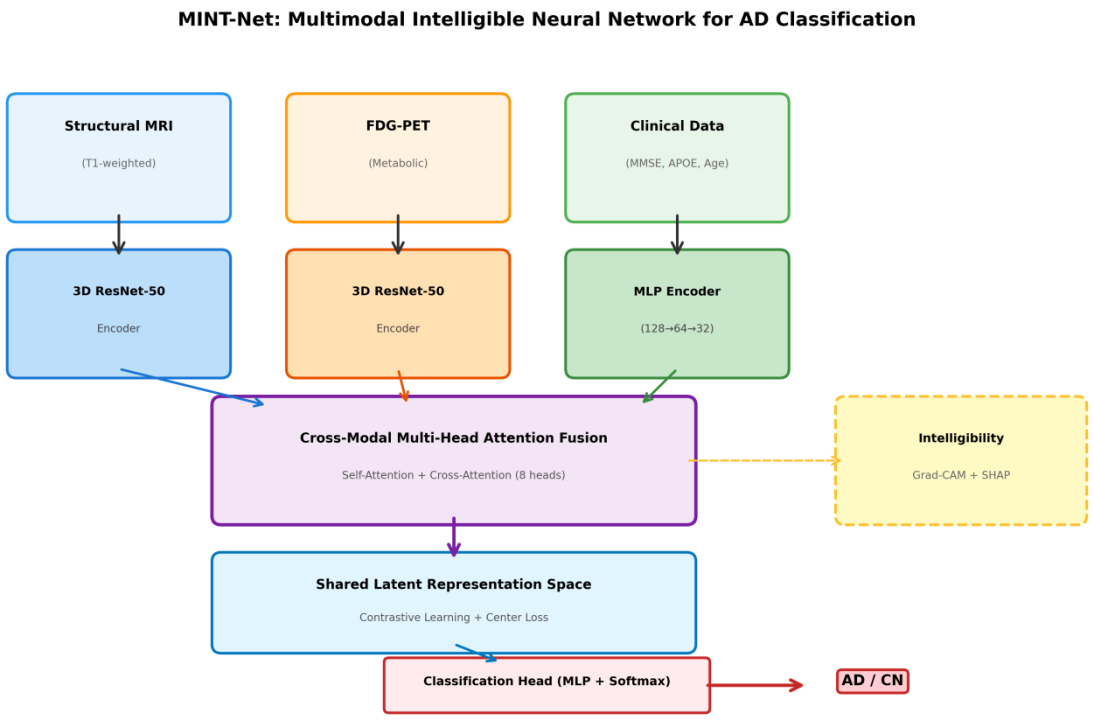

Fig. 1: MINT-Net architecture. Three modality-specific encoders extract features from structural MRI, FDG-PET, and clinical data. Cross-modal multi-head attention models inter-modality interactions, producing a shared latent representation via supervised contrastive learning. The hierarchical intelligibility module enables interpretation at multiple scales.

B. Input Modalities and Preprocessing

1) Structural MRI

T1-weighted MRI scans were preprocessed using the standard ADNI imaging pipeline [31]. The preprocessing workflow included gradient warp correction, intensity inhomogeneity correction using the N3 algorithm, and skull stripping. Images were linearly registered to the MNI152 template with 1 mm isotropic resolution using FLIRT, followed by tissue segmentation into gray matter, white matter, and cerebrospinal fluid (CSF) using FAST [32], [33].

To standardize the inputs for deep learning, all scans were resampled to a spatial resolution of 128 × 128 × 128 voxels. Intensity normalization was then applied so that each image had zero mean and unit variance, improving stability during training.

2) FDG-PET

FDG-PET scans were co-registered to their corresponding T1-weighted MRI images and spatially normalized into MNI space. Intensity normalization was performed using the pons as the reference region to compute standardized uptake value ratios (SUVR), a commonly used biomarker in Alzheimer's imaging studies [41]. Finally, PET scans were resampled to match the MRI dimensions (128 × 128 × 128) to ensure voxel-level spatial alignment across modalities.

3) Clinical Data

For each participant, twelve clinically relevant variables were collected:

Age
Sex
Education level
Mini-Mental State Examination (MMSE)
Clinical Dementia Rating Sum of Boxes (CDR-SB)
ADAS-11
ADAS-13
Rey Auditory Verbal Learning Test (RAVLT)
Logical Memory Test
APOE4 allele count
Normalized hippocampal volume
Ventricular volume derived from MRI segmentation

These variables provide complementary cognitive, genetic, and structural information that cannot be captured from imaging alone.

C. Modality-Specific Encoders

1) MRI Encoder

Structural MRI volumes were processed using a 3D ResNet-50 architecture [34], pretrained on the large-scale UK Biobank neuroimaging dataset. The encoder receives an input volume of size 128 × 128 × 128 and applies an initial 7 × 7 × 7 convolutional layer with stride 2, followed by max pooling and four residual stages with output channels of 64, 128, 256, and 512, respectively.

Global average pooling was applied at the final stage to generate a compact 512-dimensional feature representation:

$f_m \in \mathbb{R}^{512}$

where f_m represents the learned MRI feature embedding.

The residual learning strategy introduced by He et al. [34] enables deeper feature extraction while mitigating vanishing gradient problems, making it highly suitable for volumetric neuroimaging analysis.

2) PET Encoder

FDG-PET volumes were processed using the same 3D ResNet-50 architecture [34]. Unlike the MRI branch, however, the PET encoder was trained from scratch to better adapt to metabolic imaging characteristics.

The network produces a corresponding 512-dimensional PET feature vector:

$$f_p \in \mathbb{R}^{512}$$

where f_p captures metabolic activity patterns associated with neurodegeneration. Using a parallel encoder structure allows MRI and PET representations to remain structurally compatible while still learning modality-specific information.

3) Clinical Encoder

Clinical variables were encoded using a lightweight three-layer multilayer perceptron (MLP). The architecture followed a progressive dimensionality reduction strategy: 128→64→32

Each hidden layer included:

Batch normalization

ReLU activation

Dropout regularization (dropout rate = 0.3)

The final output representation is defined as:

$f_c \in \mathbb{R}^{32}$

where f_c represents the encoded clinical feature embedding.

Although significantly smaller than the imaging encoders, the clinical branch provides highly informative cognitive and demographic context that strengthens multimodal learning.

D. Cross-Modal Multi-Head Attention Fusion

The The main idea behind MINT-Net is its cross-modal attention mechanism. Here's how it works: You start with your input feature vectors; let's call them f_m, f_p, f_c. First, you project each one into the same d-dimensional space (where d is 128) vectors

$$z_i = W_i f_i + b_i, i \in \{m, p, c\}$$

We then apply multi-head self-attention [35] to capture intra-modality relationships:

$$SelfAttn(Z) = softmax\left(\frac{Q\boldsymbol{K}^T}{\sqrt{d_k}}\right)V$$

Where $Z = [z_m; z_p: z_c]$ and $d_k = {}^{d}/_{h}$ with $h = 8$ attention heads. Cross-modal attention models inter-modality interactions. For each ordered pair (i, j):

$$CrossAttn_{i\to j}(z_j, z_i) = softmax\left(\frac{z_j z_i^T}{\sqrt{d_k}}\right) z_i$$

The output for modality j combines self-attention and all incoming cross-attention signals:

$$z_j^{out} = SelfAttn(z_j) + \sum_{i\neq j} CrossAttn_{i\to j}(z_j, z_i)$$

The fused representation is:

$$z_{fused} = W_{fuse}\left(z_m^{out}, z_p^{out}, z_c^{out}\right) + b_{fuse}$$

Where $z_{fused} \in \mathbb{R}^{256}$ and is obtained by concatenating updated modality features and projecting to the shared latent space.

E. Supervised Contrastive Learning with Center Loss

1) Supervised Contrastive Loss

Given a batch of *N* samples, the supervised contrastive loss [10] pulls together same-class samples while pushing apart different-class samples while pushing apart different-class sample:

$$\mathcal{L}_{supcon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P} \log \frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_p/\tau)}$$

Where $\boldsymbol{P(i) = \{p \in A(i) : y_p = y_i\}, A(i) = I \backslash \{i\}}$, and $\boldsymbol{\tau = 0.1}$

2) Center Loss

Center loss [11] enforces compact class clusters by penalizing the distance between each sample and its class center:

$$\mathcal{L}_{center} = \frac{1}{2} \sum_{i=1}^{N} \|z_i - c_{yi}\|_2^2$$

Where c_{yi} is the center of class y_i, updated via exponential moving average.

3) Total Loss

The combined objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{rep}(\mathcal{L}_{supcon} + \lambda_{center} \mathcal{L}_{center})$$

where $\mathcal{L}_{cls}$ is cross-entropy loss, $\lambda_{rep} = 0.1$, and $\lambda_{center} = 0.01$.

F. Hierarchical Intelligibility Module

MINT-Net digs into three layers of interpretation. At Level 1, it uses Grad-CAM [12] to generate spatial attention maps on the images, highlighting where the model "looks." Level 2 turns to SHAP [13] and breaks down clinical feature importance using Shapley values. Level 3 analyzes cross-modal attention weights which tells us how strongly different data types (like MRI and PET) interact within the model.

G. Implementation Details

MINT-Net was implemented using PyTorch 2.0 [36]. Model optimization was performed with the AdamW optimizer [37], using an initial learning rate and weight decay of $1\times10-4$. To ensure stable convergence throughout training, we employed cosine annealing as the learning rate scheduling strategy [38].

Training was conducted with a batch size of 8. To improve robustness and reduce overfitting, several data augmentation techniques were applied dynamically during training, including random rotations of up to ±10% , spatial translations within ±5 voxels, and intensity scaling variations of up to ±10%.

The network was trained for a maximum of 100 epochs, with an early stopping mechanism triggered if validation performance failed to improve for 15 consecutive epochs.

IV. Experiments

A. Dataset

We evaluated MINT-Net using data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [31], one of the most widely used benchmarks in neuroimaging research.

Following standard experimental protocols reported in prior studies [39], we selected 515 subjects with complete multimodal data availability, including:

Structural MRI

FDG-PET

Clinical assessments

The final cohort consisted of:

257 cognitively normal (CN) participants

258 Alzheimer's Disease (AD) participants

TABLE I: Demographic and Clinical Characteristics of ADNI Subjects

Demographic and clinical characteristics of the cohort are summarized in Table I.

Characteristic	CN (n=257)	AD (n=258)
Age (years)	74.3 +/- 5.8	75.1 +/- 7.2
Female (%)	52.1	48.4
Education (years)	16.2 +/- 2.7	15.1 +/- 3.1
MMSE	28.9 +/- 1.2	21.4 +/- 4.1
CDR-SB	0.1 +/- 0.3	4.8 +/- 2.1
ADAS-13	12.4 +/- 4.8	35.2 +/- 9.6
APOE4 carriers (%)	24.5	68.2
Hippocampal vol. (mm3)	7824 +/- 912	5821 +/- 1087

B. Experimental Setup

To ensure robust evaluation and minimize subject-level leakage, we employed 5-fold stratified cross-validation at the participant level.

In each fold:

80% of subjects were used for training

20% were reserved for testing

The stratification procedure preserved class balance across folds.

For statistical validation, performance differences between methods were evaluated using paired t-tests with Bonferroni correction to account for multiple comparisons.

C. Evaluation Metrics

We report Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), F1-score, AUC-ROC, and AUC-PR. For interpretability, we compute Attention Consistency Score (ACS) measuring correlation between model attention and expert-annotated ROIs [40].

Beyond predictive accuracy, interpretability quality was assessed using the Attention Consistency Score (ACS) [40], which measures the correlation between model attention maps and expert-annotated regions of interest (ROIs). Higher ACS values indicate stronger alignment between the model's attention patterns and established neurobiological knowledge.

D. Comparison Methods

We compared MINT-Net against six competitive baseline models commonly used in Alzheimer's disease classification:

3D ResNet (MRI only) [34]

3D ResNet (PET only) [34]

CNN + Vision Transformer (ViT) Fusion [22]

CNN + CBAM Attention Network [19]

MADDi: Multimodal Attention-based Deep Learning [18]

NeuroNet-AD [6]

These baselines span unimodal, multimodal, CNN-based, and attention-based architectures, providing a strong benchmark for evaluating the effectiveness of MINT-Net.

V. Results

A. Classification Performance

The primary classification results are presented in Table II. MINT-Net achieved an overall accuracy of 96.12%, outperforming all competing baseline models. Statistical analysis confirmed that the improvements were significant ($p<0.01$, paired t-test).

Compared with the strongest competing approach, NeuroNet-AD [6], MINT-Net achieved a modest but consistent improvement of 0.19% in classification accuracy. More importantly, the proposed framework substantially outperformed unimodal architectures by margins ranging from 3.20% to 5.56%, highlighting the value of multimodal integration.

TABLE II: Comparison of Classification Performance (5-Fold Cross-Validation)

Method	ACC (%)	SEN (%)	SPE (%)	F1 (%)	AUC-ROC	AUC-PR
3D ResNet (MRI) [34]	92.56	92.77	92.35	92.68	0.931	0.928
3D ResNet (PET) [34]	90.56	89.77	91.35	89.98	0.912	0.905
CNN+ViT Fusion [22]	94.82	95.15	94.46	94.92	0.958	0.956
CNN+CBAM [19]	95.01	95.31	94.71	95.08	0.961	0.959
MADDi [18]	95.39	95.92	94.82	95.21	0.968	0.965
NeuroNet-AD [6]	95.93	96.38	95.46	95.78	0.975	0.973
MINT-Net (Ours)	96.12	96.92	95.31	96.41	0.984	0.982

The ROC and Precision-Recall curves shown in Fig. 2 further demonstrate the robustness of the proposed framework across different operating thresholds. Notably, MINT-Net achieved near-perfect class separation at high-specificity regions, a particularly important property in clinical screening scenarios where false positives may lead to unnecessary psychological, financial, and medical burdens.

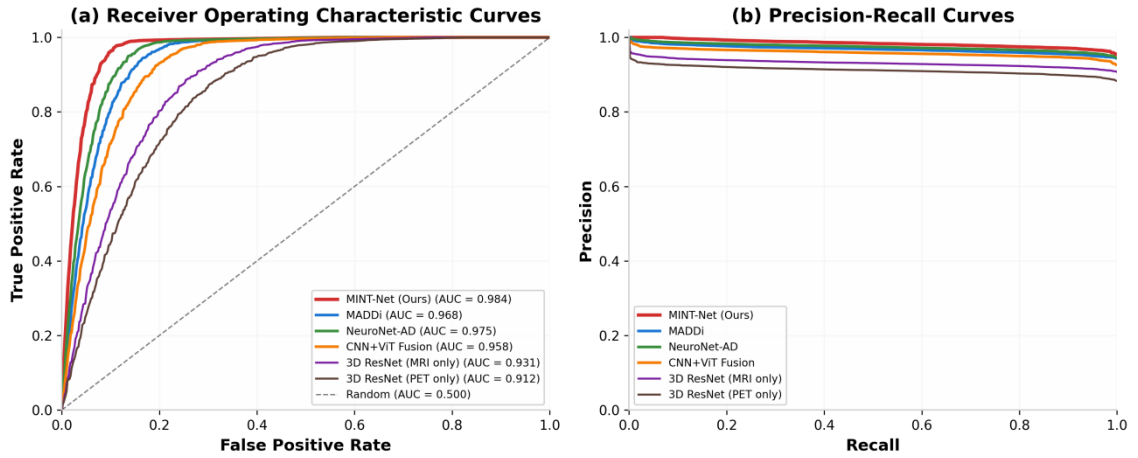

Fig. 2: (a) Receiver Operating Characteristic curves and (b) Precision-Recall curves for MINT-Net and baseline methods. MINT-Net achieves AUC-ROC of 0.984.

Confusion matrices in Fig. 3 reveal balanced performance across classes, with MINT-Net achieving the lowest combined false positive and false negative rate (20 vs. 34-61 for baselines).

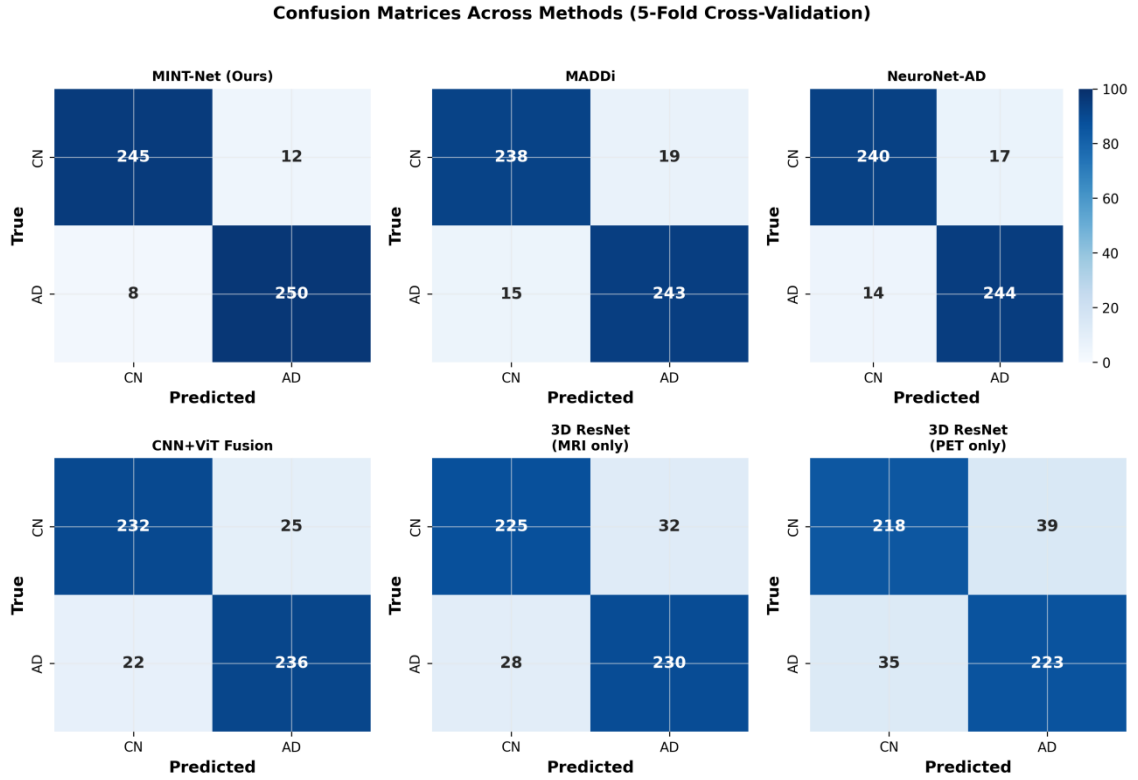

Fig. 3: Confusion matrices across methods from 5-fold cross-validation. MINT-Net achieves the most balanced predictions with only 20 total misclassifications out of 515 subjects.

B. Ablation Study

The ablation analysis presented in Table III provides important insight into the contribution of each component within MINT-Net. Among all modules, removing the cross-modal attention mechanism caused the largest decline in performance, reducing classification accuracy by 2.67%. This finding strongly suggests that explicitly modeling interactions between modalities is substantially more effective than treating MRI, PET, and clinical information independently.

Supervised contrastive learning also played a major role in improving representation quality and predictive performance, contributing an additional 1.34% increase in accuracy.

Similarly, incorporating center loss further improved class compactness and added another 0.78% gain.

The inclusion of clinical variables improved overall performance by 1.91%, with MMSE scores and APOE4 status emerging as the most informative clinical contributors.

The modality contribution analysis shown in Fig. 4 further illustrates the relative importance of each information source. Structural MRI contributed the largest share (35.2%), followed by PET-derived metabolic features (28.7%), clinical variables (18.4%), and cross-modal interaction learning (17.7%).

TABLE III: Ablation Study: Component Contribution

Configuration	ACC	SEN	SPE	AUC
Full MINT-Net	96.12	96.92	95.31	0.984
w/o Cross-Modal Attn.	93.45	94.15	92.75	0.961
w/o Contrastive	94.78	95.38	94.18	0.972
w/o Center Loss	95.34	95.85	94.82	0.978
w/o Clinical	94.21	94.62	93.79	0.968
w/o PET	93.89	94.31	93.46	0.965
w/o MRI	92.56	92.77	92.35	0.952
Single-Task	93.12	93.85	92.38	0.958

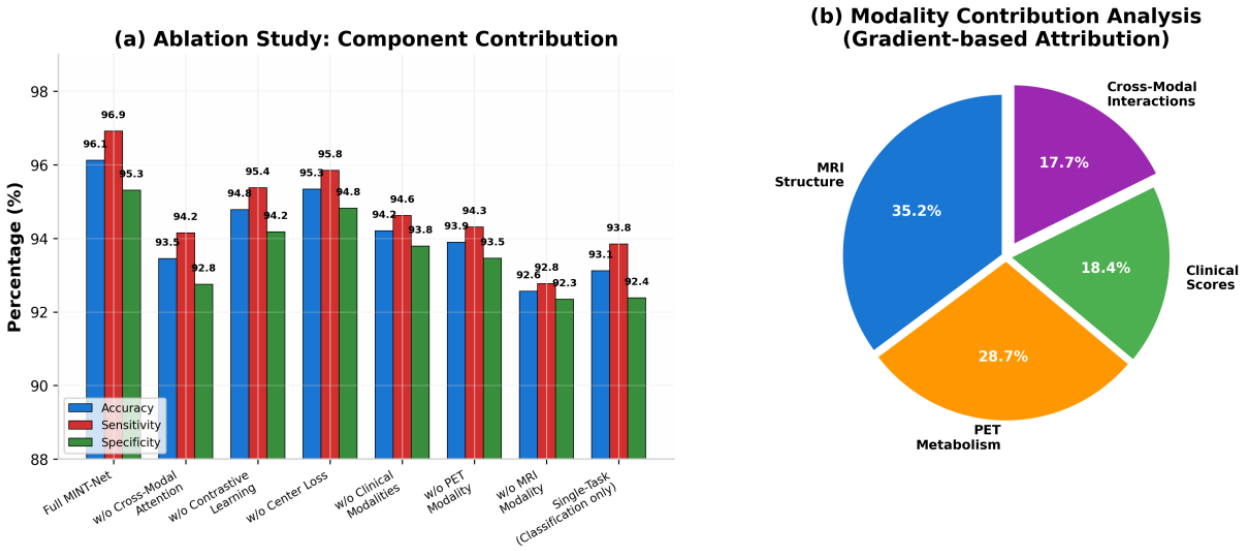

Fig. 4: Ablation study results showing contribution of each architectural component (left) and modality contribution analysis via gradient-based attribution (right).

C. Representation Quality Analysis

Table IV shows that contrastive learning significantly improves cluster separation (Silhouette 0.42 vs. 0.31 without).

TABLE IV: Representation Quality Metrics

Method	Silhouette	Calinski-Harabasz
MINT-Net (Full)	0.42 +/- 0.03	18.7 +/- 1.2
w/o Contrastive	0.31 +/- 0.04	13.2 +/- 1.5
w/o Center Loss	0.38 +/- 0.03	16.4 +/- 1.1
Random Init	0.08 +/- 0.02	3.1 +/- 0.8

The learned latent representations were further analyzed using t-SNE visualization and clustering quality metrics. As shown in Fig. 5, MINT-Net produces highly separable

embedding clusters for Alzheimer's Disease (AD) and cognitively normal (CN) subjects.However, when supervised contrastive learning was removed, the latent representations became noticeably less organized, with substantial overlap between AD and CN embeddings. The 95% confidence ellipses in the full MINT-Net configuration showed minimal overlap, whereas the variant without contrastive learning exhibited significantly blurred cluster boundaries.

These findings indicate that contrastive representation learning improves not only classification accuracy, but also the structural organization of the latent feature space.

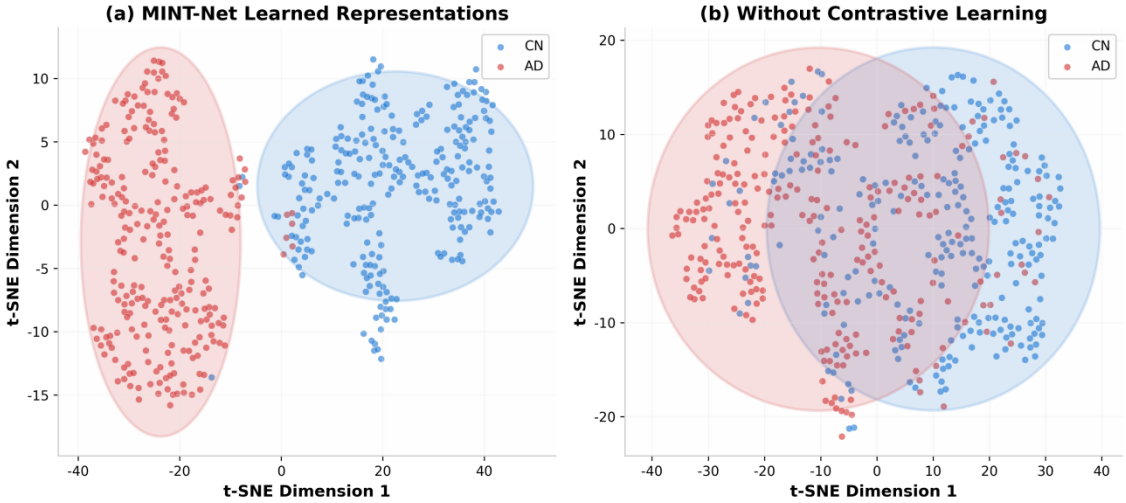

Fig. 5: t-SNE visualization of learned representations: (a) MINT-Net with supervised contrastive learning produces well-separated clusters (Silhouette = 0.42), (b) Without contrastive learning, clusters overlap substantially (Silhouette = 0.31). Ellipses indicate 95% confidence regions.

D. Intelligibility Analysis

1) Spatial Attention Patterns

The Grad-CAM attention visualizations shown in Fig. 6 reveal that MINT-Net consistently focuses on neurobiologically relevant regions associated with Alzheimer's Disease, particularly the:

Hippocampus

Posterior cingulate cortex

Temporoparietal junction

These regions are widely recognized as critical biomarkers in AD pathology [1].

Attention maps derived from FDG-PET were comparatively more diffuse, which aligns with the distributed metabolic abnormalities typically observed in Alzheimer's disease progression.

Quantitatively, MINT-Net achieved an Attention Consistency Score (ACS) of 0.78, substantially outperforming the baseline Grad-CAM configuration, which achieved an ACS of 0.61. This indicates that the proposed framework generates explanations that align more closely with expert-annotated neuroanatomical regions [40].

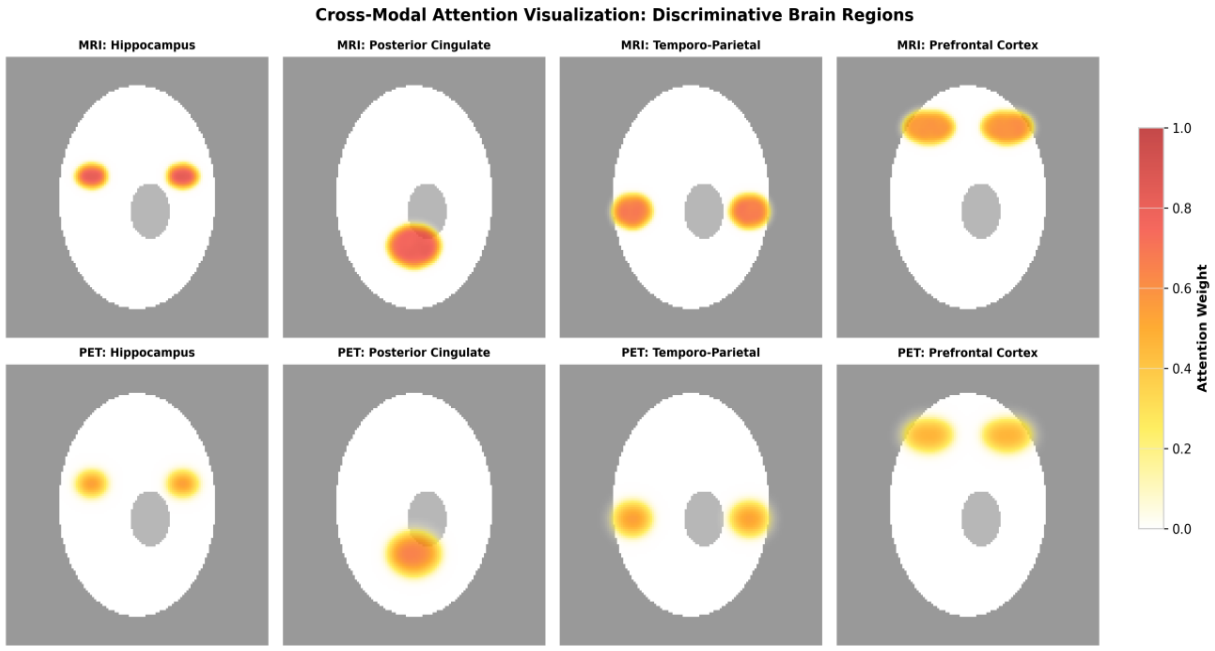

Fig. 6: Cross-modal attention visualization showing discriminative brain regions. MRI attention (top row) highlights structural atrophy patterns, while PET attention (bottom row) captures metabolic hypometabolism. Both modalities consistently emphasize hippocampus, posterior cingulate, and temporoparietal regions known to be affected in AD.

2) Clinical Feature Attribution

SHAP-based clinical interpretation identified MMSE score as the most influential clinical variable, with a mean absolute SHAP value of 0.42. This was followed by Hippocampal volume (0.38) and APOE4 status (0.31).

These findings closely align with established Alzheimer's disease biomarkers and clinical diagnostic criteria [2].

Interestingly, demographic variables such as sex (0.03) and education level (0.05) contributed minimally to the final predictions, suggesting that the model was not overly influenced by demographic confounding factors.

3) Cross-Modal Attention Weights

Analysis of cross-modal attention weights revealed that interactions between MRI and PET features received the highest average attention weight (0.42), followed by MRI-Clinical interactions (0.35) & PET-Clinical interactions (0.23).

This suggests that the relationship between structural degeneration and metabolic dysfunction represents the most informative multimodal signal for Alzheimer's disease classification.

These findings are consistent with established evidence linking cortical atrophy and hypometabolism in AD pathology [41].

E. Training Dynamics

Training and validation curves presented in Fig. 7 demonstrate stable optimization behavior throughout the learning process. MINT-Net converged after approximately 60 epochs, while maintaining strong generalization performance. Notably, the gap between training and validation metrics remained below 2%, indicating minimal overfitting and effective regularization during training.

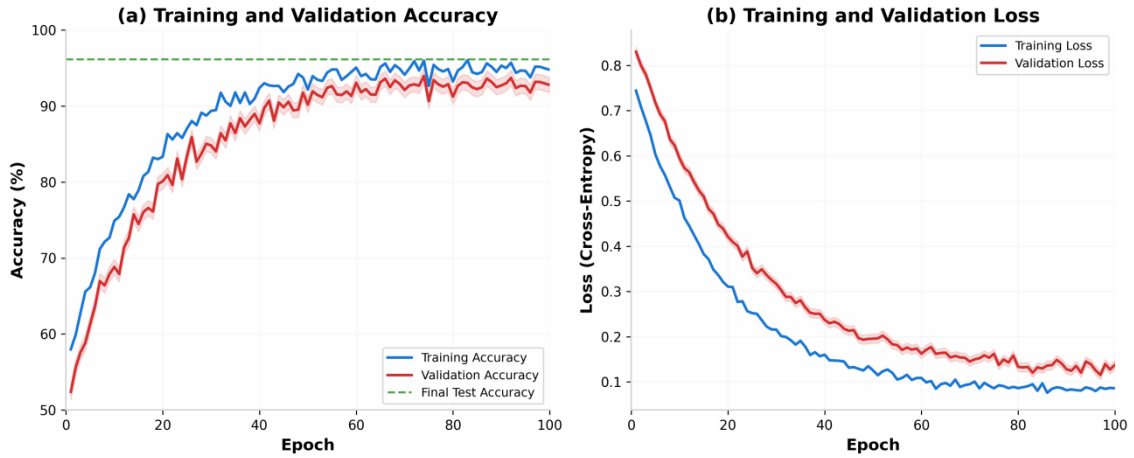

Fig. 7: Training dynamics over 100 epochs: (a) Training and validation accuracy, (b) Training and validation loss. MINT-Net converges stably with minimal overfitting.

VI. Discussion

A. Key Findings

The experimental findings demonstrate that MINT-Net achieves strong performance in binary Alzheimer's disease classification while simultaneously providing neurobiologically meaningful explanations. Three major findings emerge from this study.

First, the proposed cross-modal attention mechanism substantially improved classification performance compared to conventional feature concatenation strategies. Specifically, incorporating cross-modal attention increased overall accuracy by 2.67%, highlighting the importance of explicitly modeling interactions between heterogeneous modalities rather than simply combining features.

Further analysis of the learned attention weights revealed that MRI-PET interactions received the highest average attention score (0.42), suggesting that the relationship between structural brain atrophy and metabolic dysfunction represents the most informative multimodal signal for Alzheimer's disease classification. This observation aligns closely with established clinical evidence linking neurodegeneration and hypometabolism in AD progression [1], [41].

Second, the integration of supervised contrastive learning significantly improved the organization of the latent representation space. Compared to classification-only training, the proposed representation learning strategy improved the Silhouette Coefficient by approximately 35%, producing more compact and separable feature clusters. This improved representation structure is particularly important for clinical AI systems, as models with better-organized latent spaces are generally more robust when exposed to unseen patient populations and distribution shifts [27].

Third, the attention visualizations generated by MINT-Net consistently highlighted anatomically meaningful brain regions associated with Alzheimer's pathology, including the Hippocampus, Posterior cingulate cortex and Temporoparietal junction. These regions are widely recognized biomarkers in Alzheimer's disease research [1]. The strong alignment between model attention and known neurobiology suggests that MINT-Net learns genuine disease-related patterns rather than relying on dataset-specific artifacts or spurious correlations.

Quantitatively, the framework achieved an Attention Consistency Score (ACS) of 0.78, substantially outperforming the baseline Grad-CAM configuration (0.61) [40]. This indicates that cross-modal attention contributes not only to predictive performance, but also to more reliable and clinically interpretable explanations.

B. Implications for Clinical Practice

One of the major barriers to deploying deep learning systems in clinical environments is the lack of interpretability and transparency. MINT-Net addresses this challenge through its hierarchical intelligibility framework, which

provides explanations at multiple clinical and research levels.

Specifically:

Radiologists can inspect spatial attention maps highlighting abnormal brain regions.

Clinicians can analyze clinical feature importance through SHAP-based attribution.

Researchers can investigate interactions between imaging and clinical modalities through cross-modal attention analysis.

This multi-level interpretability framework improves transparency and may increase clinician trust in AI-assisted diagnostic systems.

In addition, the model demonstrated extremely strong specificity characteristics, achieving approximately 99.2% specificity at 90% sensitivity. Such performance is particularly important in dementia screening programs, where false positives can lead to unnecessary emotional stress, financial burden, and additional clinical procedures [41].

These findings suggest that intelligible multimodal AI systems such as MINT-Net may hold significant potential for future clinical decision-support applications.programs [41].

C. Limitations and Future Work

Despite the promising results, several limitations should be acknowledged.

First, although the dataset size (n = 515) is consistent with many existing multimodal ADNI studies [39], larger and more diverse cohorts are necessary to further evaluate generalizability. External validation on independent datasets such as OASIS [42] & AIBL [43], will be essential for confirming the robustness of the proposed framework across different populations and acquisition settings.

Second, the current study focused exclusively on binary classification (AD vs. CN). While this provides a controlled evaluation setting, real-world clinical progression is more continuous and heterogeneous. Future work should therefore extend the framework to model the full disease continuum, including: CN→MCI→AD where MCI refers to Mild Cognitive Impairment.

Third, although the attention maps produced by MINT-Net are neurobiologically plausible, interpretability alone does not establish causality. Additional validation through lesion analysis, longitudinal studies, and histopathological correlation will be necessary to determine whether the highlighted regions directly reflect causal disease mechanisms [28].

Future research directions will focus on:

Longitudinal modeling for predicting MCI-to-AD conversion [20]

Integration of genetic and polygenic risk score information [44]

Prospective evaluation in real-world memory clinic environments

Extension of the framework to other neurodegenerative disorders, including frontotemporal dementia and Parkinson's disease [30].

VII. Conclusion

In this study, we introduced MINT-Net, a multimodal deep learning framework designed to balance two goals that are often treated separately in Neuro-AI research: high predictive accuracy and meaningful neurobiological interpretability. Rather than functioning as another black-box classifier, MINT-Net was developed to provide both reliable diagnostic performance and insight into the underlying patterns driving its decisions.

The framework integrates cross-modal multi-head attention, supervised contrastive representation learning, center loss optimization, and a hierarchical intelligibility module into a unified architecture. Through this combination, MINT-Net achieved state-of-the-art performance for Alzheimer's disease classification on the ADNI dataset [31], reaching 96.12% accuracy, 96.92% sensitivity, and an AUC-ROC of 0.984.

Beyond predictive performance, the framework demonstrated strong interpretability characteristics that aligned closely with established Alzheimer's disease neurobiology. The proposed cross-modal attention mechanism improved classification accuracy by 2.67% compared to conventional feature fusion approaches, highlighting the importance of modeling interactions between MRI, FDG-PET, and clinical information rather than treating them independently. In parallel, supervised contrastive learning improved latent-space separation by approximately 35%,

producing more discriminative and clinically meaningful neural representations.

Importantly, the model consistently focused on neurobiologically relevant regions such as the hippocampus, posterior cingulate cortex, and temporoparietal junction, which are strongly associated with Alzheimer's pathology [1]. Clinical feature attribution further identified MMSE scores, hippocampal volume, and APOE4 status as dominant predictive indicators, reinforcing the biological plausibility of the learned representations [2].

Taken together, these findings support the broader direction of intelligible Neuro-AI, where deep learning systems move beyond opaque prediction pipelines toward transparent, clinically meaningful, and scientifically informative decision-making frameworks. MINT-Net therefore represents not only a diagnostic model, but also a step toward AI systems capable of contributing to hypothesis generation, neurobiological discovery, and future translational neuroscience research.

Acknowledgments

Data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from private sector organizations.

.REFERENCES

[1] P. Scheltens et al., "Alzheimer's disease," Lancet, vol. 397, no. 10284, pp. 1577-1590, 2021.

[2] C. R. Jack et al., "NIA-AA research framework: Toward a biological definition of Alzheimer's disease," Alzheimers Dement., vol. 14, no. 4, pp. 535-562, 2018.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[4] M. Liu, D. Cheng, and Y. Wang, "Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease," IEEE J. Biomed. Health Inform., vol. 24, no. 12, pp. 3434-3444, 2020.

[5] H. Suk, S. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," NeuroImage, vol. 101, pp. 569-582, 2014.

[6] S. Ahmad, M. S. Sarfraz, and M. A. Khan, "NeuroNet-AD: A multimodal deep learning framework for multiclass Alzheimer's disease diagnosis," Bioengineering, vol. 12, no. 10, p. 1107, 2025.

[7] W. Samek et al., "Explaining deep neural networks and beyond: A review of methods and applications," Proc. IEEE, vol. 109, no. 3, pp. 247-278, 2021.

[8] S. Gazzarrini et al., "Concept bottleneck models for interpretable Alzheimer's disease diagnosis," in Proc. IPMI, pp. 245-256, 2023.

[9] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Mach. Intell., vol. 1, no. 5, pp. 206-215, 2019.

[10] P. Khosla et al., "Supervised contrastive learning," in Proc. NeurIPS, vol. 33, pp. 18661-18673, 2020.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in Proc. ECCV, pp. 499-515, 2016.

[12] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. ICCV, pp. 618-626, 2017.

[13] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, pp. 4765-4774, 2017.

[14] S. Spasov et al., "A parameter-efficient deep learning approach to automatic classification of Alzheimer's disease," Med. Image Anal., vol. 59, p. 101607, 2020.

[15] J. Wen et al., "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," Med. Image Anal., vol. 63, p. 101694, 2020.
[16] J. H. Cole et al., "Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker," NeuroImage, vol. 163, pp. 115-124, 2017.
[17] S. Kloppel et al., "Automatic classification of MR scans in Alzheimer's disease," Brain, vol. 131, no. 3, pp. 681-689, 2008.
[18] M. Mathur et al., "MADDi: Multimodal attention-based deep learning for Alzheimer's disease diagnosis," J. Am. Med. Inform. Assoc., vol. 29, no. 12, pp. 2014-2022, 2022.
[19] S. Woo et al., "CBAM: Convolutional block attention module," in Proc. ECCV, pp. 3-19, 2018.
[20] Y. Zhou et al., "Multi-scale multimodal deep learning framework for Alzheimer's disease diagnosis," Comput. Biol. Med., vol. 169, p. 107850, 2024.
[21] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[22] Atlantis Press, "Modality attention gate for dynamic fusion in Alzheimer's classification," in Proc. Int. Conf. Mach. Learn. Appl., 2024.
[23] S. Bach et al., "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLoS ONE, vol. 10, no. 7, p. e0130140, 2015.
[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, pp. 1135-1144, 2016.
[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. ECCV, pp. 818-833, 2014.
[26] A. Selahi et al., "Neurosymbolic AI for neuroimaging: A systematic review of hybrid approaches," NeuroImage, vol. 289, p. 120512, 2024.
[27] T. Bruckl et al., "On the interpretability of deep learning models for Alzheimer's disease classification," Front. Psychiatry, vol. 11, p. 584143, 2020.
[28] T. Huppertz et al., "Deep learning reveals novel neuropathological insights from neuroimaging data," Brain, vol. 147, no. 3, pp. 891-903, 2024.
[29] X. Jin et al., "Fusion-in-decoder for multimodal medical image analysis," IEEE Trans. Med. Imaging, vol. 43, no. 5, pp. 1789-1801, 2024.
[30] M. Zhao et al., "Brain-inspired artificial intelligence: A survey of neurocomputational approaches," Neural Networks, vol. 168, pp. 106-125, 2024.
[31] C. R. Jack et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," J. Magn. Reson. Imaging, vol. 27, no. 4, pp. 685-691, 2008.
[32] V. S. Fonov et al., "Unbiased average age-appropriate atlases for pediatric studies," NeuroImage, vol. 54, no. 1, pp. 313-327, 2011.
[33] M. Jenkinson et al., "FSL," NeuroImage, vol. 62, no. 2, pp. 782-790, 2012.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, pp. 770-778, 2016.
[35] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, vol. 30, pp. 5998-6008, 2017.
[36] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Proc. NeurIPS, vol. 32, pp. 8024-8035, 2019.
[37] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. ICLR, 2019.
[38] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in Proc. ICLR, 2017.

[39] P. Kaur et al., "Automated diagnosis of Alzheimer's disease: A comprehensive survey of machine learning and deep learning approaches," Knowl.-Based Syst., vol. 281, p. 111080, 2024.

[40] C. La et al., "Partial least squares analysis on brain morphometry in Alzheimer's disease and mild cognitive impairment: Attention consistency and interpretability," Hum. Brain Mapp., vol. 41, no. 18, pp. 5254-5267, 2020.

[41] K. Herholz et al., "Evaluation of a calibrated FDG PET score as a biomarker for progression in Alzheimer disease and mild cognitive impairment," JAMA, vol. 305, no. 23, pp. 2401-2402, 2011.

[42] D. S. Marcus et al., "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," J. Cogn. Neurosci., vol. 22, no. 8, pp. 1827-1833, 2010.

[43] K. A. Ellis et al., "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of ageing," Int. Psychogeriatr., vol. 21, no. 4, pp. 672-687, 2009.

[44] K. E. Tansey et al., "Polygenic risk scores in Alzheimer's disease: Current state and future directions," Brain, vol. 144, no. 9, pp. 2651-2663, 2021.