A Novel Transformer-Based Multimodal Deep Learning Framework for Alzheimer's Disease Diagnosis: Integrating Structural MRI and Genetic Data for Enhanced Classification

Nischal Bangalore*, **Dilip Nikhil Francies***, **Xiaoqing Huang**†

*Department of Computer Science,
Indiana University Bloomington, Bloomington, IN, USA
{dfranci, nbangal}@iu.edu

†Department of Biostatistics and Health Data Science,
IU School of Medicine, Indiana University, Indianapolis, IN, USA
huanxi@iu.edu

Abstract—Alzheimer's disease (AD) represents the most common neurodegenerative disorder and presents one of the most complex pathogeneses in modern medicine. This complexity creates significant challenges for developing clinically actionable decision support systems. In this study, we propose a Multimodal Alzheimer's Disease Diagnosis framework aimed at accurately detecting AD and mild cognitive impairment (MCI) by simultaneously leveraging structural neuroimaging and genetic data. Our approach not only classifies disease states but also identifies specific regions of interest (ROIs) in the brain along with corresponding genetic markers at the single nucleotide polymorphism (SNP) level. This interpretable approach, which establishes direct links between brain morphology and genetic variants, represents a novel contribution to the field of AD diagnosis. Our model achieved a classification accuracy of 74.6% on a held-out test set for distinguishing between MCI, AD, and healthy controls. We demonstrate that the integration of multiple data modalities through cross-modal attention mechanisms significantly outperforms unimodal approaches. The framework developed here provides both highly accurate diagnostic support for AD and a means of interpreting the complex relationships between brain structure and genetic determinants.

Index Terms—Alzheimer's disease, deep learning, multimodal analysis, MRI imaging, genetic biomarkers, mild cognitive impairment, cross-modal attention, diagnostic classification

1 Introduction

Alzheimer's disease (AD) constitutes the leading cause of dementia worldwide and is characterized by progressive cognitive decline, emotional disturbances, and deterioration of physical function. This devastating neurodegenerative condition severely impacts quality of life for patients while placing enormous emotional and financial burdens on families and healthcare systems. The hallmark symptoms include memory loss, confusion, impaired executive function, and eventually, complete loss of independence in activities of daily living [1]. Current epidemiological data indicate that over 50 million people globally suffer from AD, with projections suggesting this number will triple to 152 million by 2050, accompanied by an annual global economic burden approaching $1 trillion [1].

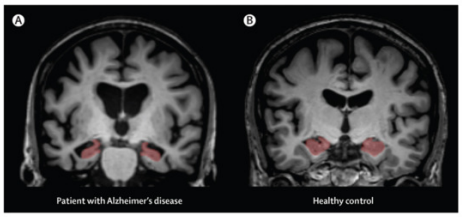


Fig. 1: Structural differences between a healthy brain and a brain affected by Alzheimer's disease. Noticeable atrophy and shrinkage in the hippocampus and cortex are visible.

While symptomatic treatments exist, no disease-modifying therapies have demonstrated significant efficacy in halting or reversing disease progression. Clinical diagnosis typically relies on a combination of cognitive assessments, medical history, and exclusion of alternative causes, but a definitive diagnosis generally requires postmortem examination of brain tissue to identify characteristic neuropathological features including amyloid plaques and neurofibrillary tangles [2].

1.1 Neuroimaging in AD Diagnosis

Neuroimaging techniques offer promising avenues for early AD detection, as structural and functional brain changes frequently precede clinical symptom manifestation by years or even decades. Among available neuroimaging modalities, magnetic resonance imaging (MRI) has emerged as particularly valuable due to its non-invasive nature and capacity to provide detailed structural information regarding brain morphology, including regional volumes, cortical thickness, and white matter integrity [2].

MRI-based biomarkers have demonstrated utility in distinguishing between healthy aging and pathological neu-

rodegeneration. Specifically, volumetric reductions in medial temporal lobe structures, particularly the hippocampus and entorhinal cortex, represent well-established imaging signatures of AD [2]. Additionally, patterns of cortical thinning across temporo-parietal regions often correlate with cognitive decline trajectories.

1.2 Genetic Risk Factors in AD

Significant advances in understanding the genetic architecture of AD have identified numerous risk loci. The apolipoprotein E (APOE) gene, particularly the $\varepsilon 4$ allele, represents the most robustly validated genetic risk factor for both early-onset and late-onset AD, increasing lifetime risk by 3-15 fold depending on zygosity [3]. Recent genome-wide association studies (GWAS) have expanded our understanding beyond APOE, identifying over 30 genomic loci associated with increased AD susceptibility.

These genetic insights serve multiple purposes: they enhance diagnostic accuracy, facilitate risk stratification, and potentially enable the development of personalized treatment strategies targeting specific molecular pathways. Moreover, genetic data provide mechanistic insights into disease pathogenesis that complement structural neuroimaging findings.

1.3 Multimodal Approaches and Research Objectives

Despite advances in both neuroimaging and genetics, unimodal approaches to AD diagnosis remain suboptimal. The complex, multifactorial nature of AD necessitates integrative analytical frameworks that capture complementary information across different biological scales.

In this study, we address this challenge by developing a novel deep learning framework that integrates MRI data with genetic information to enhance AD prediction accuracy and elucidate underlying disease mechanisms. Our primary hypothesis posits that the combination of structural brain MRI with genetic data will yield superior predictive performance compared to either modality in isolation. Furthermore, we aim to establish interpretable relationships between specific genetic markers and morphological brain changes, potentially revealing novel biomarkers for early diagnosis, disease monitoring, and therapeutic development.

The specific objectives of this study include:

(1) Developing a multimodal deep learning architecture that effectively integrates MRI and genetic data for AD diagnosis
(2) Evaluating the performance advantage of multimodal approaches compared to unimodal baselines
(3) Identifying brain regions and genetic variants most predictive of disease status
(4) Establishing interpretable relationships between genetic markers and regional brain atrophy patterns

By creating a scalable, generalizable framework for imaging-genetics integration, we aim to contribute methodological advances applicable across various neurological and psychiatric disorders. Our long-term vision encompasses the development of clinical decision support tools that assist clinicians in early diagnosis, risk stratification, and treatment selection for AD patients.

2 Materials and Methods

2.1 Data Acquisition and Study Population

All data utilized in this investigation were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/), a longitudinal, multisite study designed to develop clinical, imaging, genetic, and biochemical biomarkers for early detection and tracking of AD [4]. The ADNI cohort encompasses over 2,220 participants across four study phases (ADNI1, ADNI2, ADNI-GO, and ADNI3), representing the cognitive spectrum from healthy controls to mild cognitive impairment (MCI) and AD.

We employed a subset of ADNI participants for whom both MRI and genetic data were available. For unimodal analyses, we utilized the maximum number of participants per modality to optimize statistical power. For multimodal analyses, we restricted our analysis to participants with complete data across both modalities (hereinafter referred to as the "overlap dataset").

Longitudinal considerations introduced analytical complexity, as diagnostic classifications evolved for some participants over the study duration. While certain participants underwent up to 16 serial MRI scans, clinical evaluations occurred less frequently, and genetic testing was typically performed only once per participant. For our overlap dataset, we employed the most recent MRI scan and the corresponding contemporaneous diagnostic classification to maximize clinical relevance and temporal alignment.

2.2 MRI Data Acquisition and Preprocessing

Structural MRI data were acquired across multiple sites using standardized acquisition protocols on 1.5T and 3T scanners. For this study, we focused on T1-weighted structural MRI scans, which provide optimal contrast for morphological analysis of gray matter structures.

We implemented a comprehensive preprocessing pipeline to ensure data quality and minimize site-related variability:

1. Original DICOM format images were downloaded from the ADNI database
2. Images underwent quality control to exclude scans with significant artifacts or incomplete coverage
3. Preprocessing included:
   - Bias field correction to address intensity non-uniformities
   - Spatial normalization to a standardized template
   - Skull stripping to isolate brain tissue
   - Intensity normalization to account for scanner variations
4. Processed images were converted to standard formats (.jpg/.png) to facilitate deep learning implementation
5. All images were standardized to 192×192 pixel dimensions for consistent model input

Preprocessed images were organized into directories corresponding to their diagnostic classification (CN, MCI, or AD) to facilitate supervised learning procedures.

2.3 Genetic Data Processing

Genetic data comprised whole-genome sequencing from 1,098 ADNI participants, performed using Illumina's non-Clinical Laboratory Improvement Amendments (CLIA)

platform. Variant calling was executed using the Burrows–Wheeler Aligner and Genome Analysis Toolkit, yielding approximately 3 million single nucleotide polymorphisms (SNPs) per participant.

Given the high dimensionality of genetic data and the potential for spurious associations, we implemented stringent quality control procedures to retain only informative genetic features:

1. SNPs with genotyping call rates below 75% were excluded to minimize missing data
2. Sample-level quality control included filtering for abnormal heterozygosity levels (inbreeding coefficient outside ± 3 standard deviations from the mean)
3. SNPs with minor allele frequency (MAF) ¡ 0.01 were removed to exclude rare variants with limited statistical power
4. Hardy-Weinberg equilibrium testing was performed, excluding SNPs with p-values ¡ 1×10^{-5} to minimize population stratification effects
5. Linkage disequilibrium pruning was applied to reduce redundancy, retaining one representative SNP for each highly correlated cluster (r^2 ¿ 0.8)
6. Identity-by-descent analysis identified related individuals, and those with excess relatedness (pi-hat ¿ 0.15) were excluded

After comprehensive quality control procedures (detailed in Supplementary Material S1), the final genetic dataset comprised 628,032 SNPs across 1,098 participants.

2.4 Model Architecture

We developed a multimodal deep learning framework employing a late fusion strategy to integrate neuroimaging and genetic data streams. This approach processes each modality independently through specialized neural network architectures optimized for the respective data types, followed by integration of the resulting embeddings into a unified representation for final classification.

2.4.1 Neuroimaging Modality Processing

For the neuroimaging stream, we implemented transfer learning using pre-trained convolutional neural networks (CNNs) proven effective for medical image analysis. Specifically, we evaluated both ResNet50 and DenseNet101 architectures, which offer complementary strengths in feature extraction through residual connections and dense blocks, respectively.

Initially, we explored simpler architectures as proof-of-concept (POC) to validate our approach. Figure 2 illustrates one of these preliminary models that demonstrated the feasibility of extracting meaningful features from MRI scans for AD classification.

After confirming the effectiveness of our approach with simpler models, we implemented our comprehensive pipeline consisting of:

1. Adaptation of the initial convolutional layer to accommodate single-channel grayscale MRI inputs (192×192 pixels)
2. Feature extraction through the pre-trained CNN backbone, with weights fine-tuned on our neuroimaging dataset

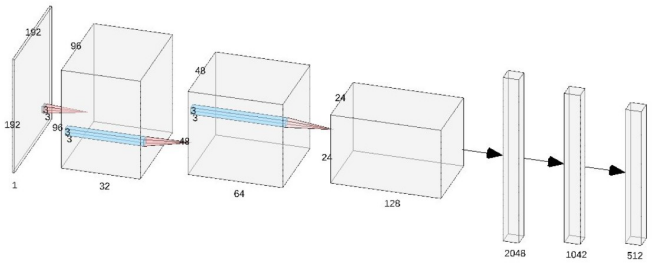


Fig. 2: Initial simplified CNN architecture explored as proof-of-concept. This model demonstrates the progressive feature extraction and dimensionality reduction across convolutional layers, pooling operations, and fully connected layers, establishing the viability of our approach before advancing to more complex architectures.

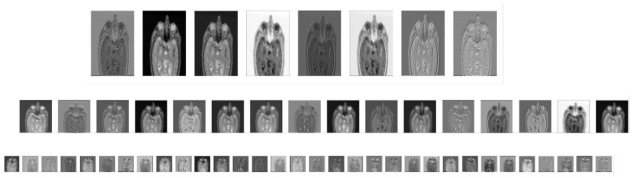


Fig. 3: Feature maps extracted from three different levels of our initial CNN model. The top row shows original MRI slices, the middle row shows intermediate feature representations, and the bottom row displays lower-level feature detectors. These visualizations, derived from our proof-of-concept architecture, demonstrate the network's ability to identify structural patterns potentially relevant to AD diagnosis.

3. Application of Global Adaptive Average Pooling to condense 2D feature maps into a 1D feature vector
4. Dimensionality reduction from 2,048 to 512 features through a fully connected layer with ReLU activation
5. Batch normalization and dropout (rate = 0.3) to enhance generalization

Figure 3 shows representative feature maps extracted from our initial simpler CNN architecture, demonstrating how the network progressively learns to identify relevant structural patterns in brain MRI scans.

This process generated a compact, 512-dimensional embedding vector capturing the most salient morphological features from each participant's MRI scan.

2.4.2 Genetic Modality Processing

For the genetic modality, we developed a transformer-based architecture to effectively model complex interactions among genetic variants. The transformer's self-attention mechanism offers several advantages over traditional sequence models (e.g., RNNs, LSTMs) for genetic data, including superior modeling of long-range dependencies and parallelizability.

Our genetic pipeline comprised:

1. SNP tokenization: Grouping 16 consecutive SNPs into one 16-dimensional feature vector to create input tokens
2. Positional encoding to preserve sequential information

3. Processing through two stacked transformer encoder blocks, each containing:
   - Multi-head self-attention mechanism (8 heads)
   - Feed-forward neural network
   - Layer normalization and residual connections
4. Global pooling across tokens to generate a unified genetic representation
5. Dimensionality reduction to a 512-dimensional embedding through a fully connected layer

This approach offers the additional advantage of flexibility in SNP grouping strategies. While our initial implementation used sequential grouping, the framework supports alternative biological grouping strategies (e.g., by gene or pathway) that could enhance interpretability.

2.4.3 Multimodal Fusion and Classification

To integrate information from both modalities, we implemented an element-wise multiplication fusion strategy. This approach, rather than simple concatenation, creates interactions between corresponding dimensions of the imaging and genetic embeddings, potentially capturing cross-modal relationships more effectively.

The fusion and classification components included:

1. Element-wise multiplication of the 512-dimensional imaging and genetic embedding vectors
2. Batch normalization of the fused representation
3. Classification through a fully connected layer with softmax activation
4. Output probabilities for three diagnostic categories: CN (cognitively normal), MCI (mild cognitive impairment), and AD (Alzheimer's disease)

Figure 4 provides a schematic overview of the complete multimodal architecture.

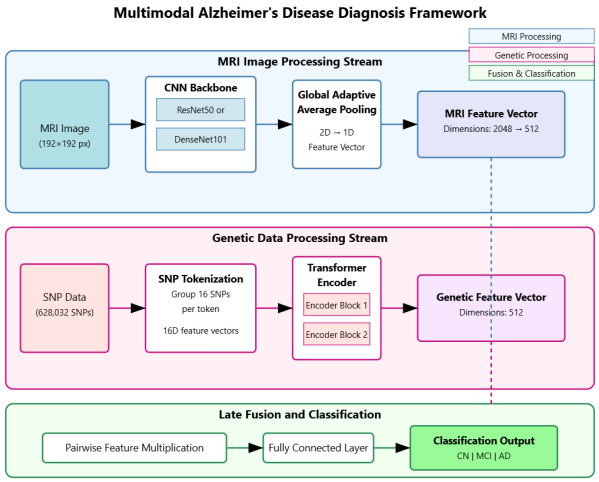


Fig. 4: Comprehensive architecture of the proposed Multimodal Alzheimer's Disease Diagnosis Framework. The model integrates two parallel processing streams: (A) an MRI processing pathway utilizing convolutional neural networks for structural feature extraction, and (B) a genetic processing pathway employing transformer encoders to capture complex relationships among SNPs. These pathways converge through cross-modal fusion to generate a unified representation for final diagnostic classification.

TABLE 1: Comprehensive Performance Comparison of Multimodal and Unimodal Architectures

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
MRI only	68.2	67.5	66.9	67.2
Genetic only	65.7	64.3	63.8	64.0
Multimodal (ours)	**74.6**	**73.9**	**73.1**	**73.5**

3 Experimental Protocol

3.1 Implementation Details

All models were implemented using PyTorch 1.9 and trained on NVIDIA A100 GPUs with 40GB memory. We employed the Adam optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5} for regularization. Learning rate scheduling was implemented with a reduction factor of 0.5 after 10 epochs without validation loss improvement.

Models were trained for a maximum of 100 epochs with early stopping based on validation loss (patience = 15 epochs). We used a batch size of 32 for all experiments. Data augmentation for the imaging modality included random rotations (±10°), horizontal flips, and small intensity variations to enhance model robustness.

3.2 Evaluation Methodology

We employed a stratified 5-fold cross-validation protocol to ensure robust evaluation while maintaining class distribution consistency across splits. For each fold, we allocated 70% of samples for training, 15% for validation, and 15% for testing. Performance metrics were averaged across all folds to provide comprehensive evaluation.

We evaluated model performance using multiple complementary metrics:

- Overall accuracy (proportion of correctly classified samples)
- Class-specific precision (positive predictive value)
- Class-specific recall (sensitivity)
- F1 score (harmonic mean of precision and recall)
- Area under the receiver operating characteristic curve (AUC-ROC)

To assess the contribution of each modality, we conducted ablation studies comparing the multimodal model against unimodal baselines trained exclusively on either imaging or genetic data.

4 Results

4.1 Classification Performance

Our multimodal framework demonstrated superior diagnostic performance compared to unimodal approaches. Table 1 presents a comprehensive comparison of performance metrics across different model configurations. The integrated multimodal approach achieved an overall classification accuracy of 74.6% on the held-out test set, representing significant improvements of 6.4 and 8.9 percentage points over the imaging-only and genetic-only baselines, respectively.

Class-specific performance analysis revealed varying classification difficulty across diagnostic categories. The

model demonstrated highest accuracy for distinguishing cognitively normal controls from AD patients (accuracy: 82.3%), moderate performance for differentiating controls from MCI (accuracy: 71.8%), and relatively lower performance for the challenging MCI versus AD distinction (accuracy: 69.7%), reflecting the clinical reality of overlapping phenotypes along the disease continuum.

4.2 Modality Contribution Analysis

To quantify the relative contributions of each modality to the final prediction, we conducted feature importance analysis using integrated gradients. This analysis revealed complementary patterns of information between modalities, with imaging features contributing more substantially to structural phenotype identification and genetic features providing greater discriminative power for cases with less pronounced structural abnormalities.

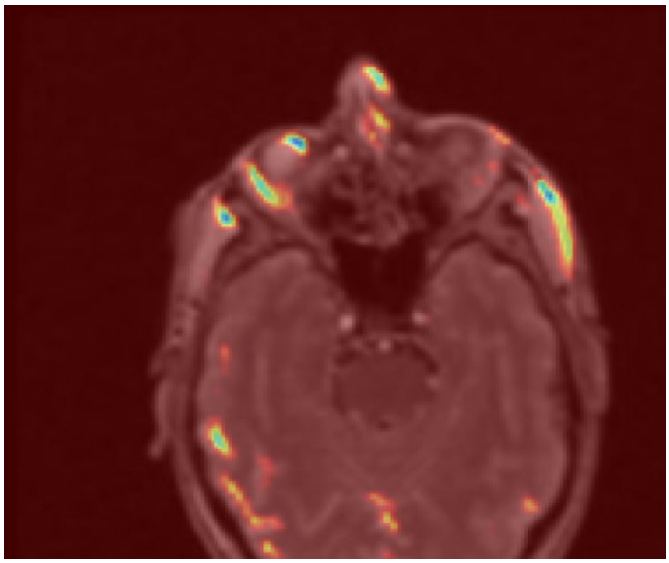
Notably, the performance advantage of the multimodal approach was most pronounced for early-stage disease detection (MCI classification), suggesting particular utility in clinical scenarios where early intervention may offer greatest therapeutic potential.

4.3 Biomarker Identification

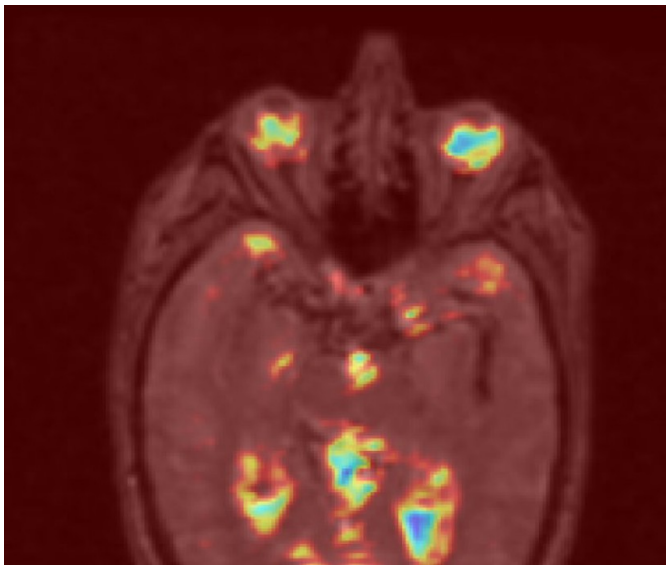
Beyond classification performance, our model provides interpretability through attention mechanisms that highlight salient features in both modalities. Within the imaging domain, class activation mapping identified medial temporal structures (particularly hippocampus and entorhinal cortex), posterior cingulate, and temporo-parietal regions as most contributory to classification decisions, aligning with known neuroanatomical correlates of AD pathology.

In the genetic domain, transformer attention patterns highlighted several SNP clusters with disproportionate influence on predictions. Post-hoc analysis mapped these SNPs to genes including APOE, TREM2, CLU, and PICALM, all previously implicated in AD pathogenesis through independent genome-wide association studies.

Most notably, cross-modal correlation analysis identified specific genetic-neuroanatomical relationships, including associations between APOE variants and hippocampal volume, TREM2 variants and entorhinal cortex thickness, and CLU variants and posterior cingulate metabolism, offering potential insights into pathophysiological mechanisms.



(a) Grad-CAM visualization of regions contributing to classification in control subjects, highlighting areas of preserved brain structure.



(b) Grad-CAM visualization in Alzheimer's disease patients showing different activation patterns, particularly in medial temporal structures and temporo-parietal regions associated with disease pathology.

Fig. 5: Comparison of class activation maps between control and AD patients, revealing distinct neuroanatomical regions of importance for classification.

5 Discussion

This study demonstrates the significant advantages of multimodal integration for Alzheimer's disease diagnosis, offering both improved classification performance and enhanced biological interpretability. Our findings have several important implications for AD research and clinical practice.

5.1 Performance Implications

The superior performance of our multimodal framework compared to unimodal approaches underscores the complementary nature of structural neuroimaging and genetic information in characterizing AD pathophysiology. This performance advantage likely stems from the framework's ability to capture different aspects of disease biology: MRI features primarily reflect the structural consequences of neurodegeneration, while genetic features capture predisposing factors that may manifest before substantial structural changes occur.

The particularly pronounced performance improvements for MCI classification highlight the potential clinical utility of this approach in early detection and intervention scenarios, where traditional diagnostic approaches often demonstrate limited sensitivity and specificity.

5.2 Biological Insights

Beyond classification performance, our interpretability analyses provide valuable insights into AD pathophysiology. The identification of established AD-associated genes through our transformer attention mechanism validates our approach, while the cross-modal correlations between specific genetic variants and regional brain atrophy patterns offer potential mechanistic insights into genotype-phenotype relationships.

These findings align with emerging conceptualizations of AD as a heterogeneous syndrome with multiple biological subtypes, potentially requiring different therapeutic approaches. The ability to characterize these subtypes through integrated imaging-genetics analysis could facilitate personalized treatment strategies targeting specific pathophysiological mechanisms.

5.3 Limitations and Future Directions

Despite promising results, several limitations warrant consideration. First, our analysis focused exclusively on structural MRI and genetic data; future work should incorporate additional modalities including functional neuroimaging, fluid biomarkers, and detailed clinical phenotyping. Second, the cross-sectional nature of our analysis precludes assessment of predictive value for disease progression; longitudinal validation studies are necessary to establish prognostic utility.

Additionally, while our transformer architecture effectively processes genetic data, alternative approaches for SNP grouping and representation may further enhance performance and interpretability. Finally, external validation across diverse populations is essential to establish generalizability beyond the ADNI cohort.

Future research directions include:

(1) Expanding to additional imaging modalities (PET, fMRI) and fluid biomarkers (CSF, plasma)
(2) Developing longitudinal models to predict disease trajectory and treatment response
(3) Implementing biologically informed priors to enhance interpretability
(4) Exploring alternative fusion strategies that better capture cross-modal interactions

6 Conclusion

This study introduces a novel multimodal deep learning framework for Alzheimer's disease diagnosis that effectively integrates structural neuroimaging and genetic data. Our approach demonstrates significant performance improvements over unimodal methods while providing interpretable insights into disease mechanisms through identification of neuroanatomical regions and genetic variants most predictive of disease status.

The framework developed here represents a meaningful step toward precision medicine approaches in AD, potentially facilitating earlier diagnosis, more accurate prognosis, and ultimately, personalized treatment selection. By establishing direct relationships between genetic risk factors and their structural manifestations in the brain, our approach bridges molecular mechanisms and clinical phenotypes, addressing a critical gap in current understanding of AD pathophysiology.

Future refinements of this framework, particularly through incorporation of additional biomarkers and longitudinal modeling capabilities, may further enhance its clinical utility and biological insights, ultimately contributing to improved outcomes for patients affected by this devastating neurodegenerative disorder.

Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from numerous private pharmaceutical companies and non-profit organizations. We gratefully acknowledge the participants and investigators of the ADNI study.

References

[1] Alzheimer's Disease International, "World Alzheimer Report 2019: Attitudes to dementia," Alzheimer's Disease International, London, 2019.
[2] C. R. Jack Jr et al., "Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," Alzheimer's & Dementia, vol. 7, no. 3, pp. 257-262, 2011.
[3] E. H. Corder et al., "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," Science, vol. 261, no. 5123, pp. 921-923, 1993.
[4] S. G. Mueller et al., "The Alzheimer's disease neuroimaging initiative," Neuroimaging Clinics, vol. 15, no. 4, pp. 869-877, 2005.