UNIVERSITY OF
BIRMINGHAM

Transfer Learning for Alzheimer's Disease Detection: Adapting Video Classification Models for MRI Scans

Rhys W. Alexander (2458177)

Final project report submitted
in partial fulfilment for the degree of
B.SCI. IN ARTIFICIAL INTELLIGENCE AND COMPUTER SCIENCE

Date: 7th April 2025
Word count: X,XXX

Project supervisor:
Dr Rickson Mesquita

Contents

1 Abstract

2 Introduction

3 Literature Review

This review synthesizes current knowledge across medical and computational domains relevant to Alzheimer's disease detection using deep learning approaches, examining AD neuroimaging biomarkers, computational approaches, and research gaps.

3.1 Alzheimer's Disease and Neuroimaging

3.1.1 Pathophysiology with Emphasis on Structural Changes

Alzheimer's disease pathophysiology follows a predictable cascade, beginning with amyloid β deposition and hyperphosphorylated tau aggregation, which precede detectable structural changes [1]. These processes ultimately manifest as progressive neurodegeneration visible on structural MRI, see figure 1. The hippocampus and entorhinal cortex are among the earliest affected regions, showing measurable atrophy years before clinical symptoms emerge. This atrophy pattern subsequently extends to temporal, parietal, and frontal cortices, correlating closely with cognitive decline [2]. Structural MRI can detect these volumetric changes with high sensitivity, providing quantitative biomarkers that reflect underlying neuronal loss.

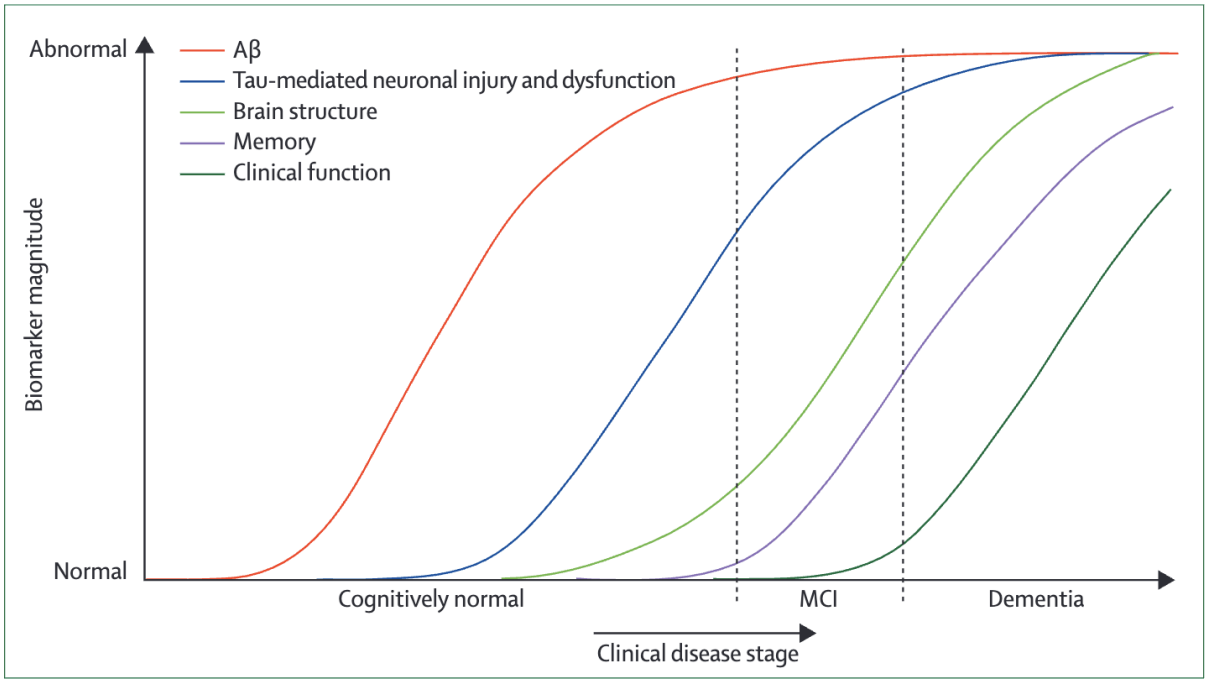

Figure 1: Temporal progression of AD biomarkers, showing the relative timeline of pathophysiological changes [1].

3.1.2 Hippocampal Atrophy as Primary Biomarker

Hippocampal atrophy represents one of the earliest and most established structural biomarkers in Alzheimer's disease progression [3]. Volume reductions follow a predictable pattern, beginning years before clinical symptoms emerge, with annual atrophy rates of 3-6% in AD compared to 1-2% in normal aging [2]. Volumetric measurements correlate strongly with cognitive decline and Braak staging of neurofibrillary pathology. Standardized quantification methods include manual tracing, automated segmentation, and shape analysis, achieving diagnostic sensitivities of 80-90% and specificities of 80-95% in distinguishing AD from healthy controls [4].

3.1.3 Additional Neuroimaging Markers

Beyond volumetric measurements, shape analysis methods capture morphological changes in brain structures [5], detecting subtle deformations missed by volume alone. Other promising markers include cortical thickness measurements [6], white matter integrity via diffusion tensor imaging, functional connectivity patterns, and metabolic alterations detectable through PET imaging [2]. These diverse markers provide complementary information that may enhance transfer learning models' diagnostic accuracy.

3.1.4 Current Clinical Diagnostic Practices and Limitations

Current AD diagnosis follows NINCDS-ADRDA criteria, integrating clinical assessment, cognitive testing, and biomarker analysis [7]. Visual assessment of neuroimaging suffers from significant inter-reader variability, with diagnostic accuracy dependent on radiologist expertise [4]. A substantial temporal gap exists between initial pathological changes and clinical manifestation, complicating early intervention [8]. Additionally, clinical diagnostic accuracy ranges from 65-96%, with lower precision in early disease stages when intervention would be most beneficial [9].

3.1.5 Role of Structural MRI and T1-weighted Imaging in Diagnosis

Structural MRI provides objective evidence of neurodegeneration that complements clinical assessment, with hippocampal atrophy serving as a primary biomarker [7]. MRI"s advantages include non-invasiveness compared to CSF sampling, absence of radiation exposure unlike PET, wider availability, and lower cost [2]. However, visual assessment suffers from inter-reader variability and limited sensitivity to subtle changes, with accuracy heavily dependent on radiologist expertise [9]. These limitations underscore the need for quantitative, automated analysis approaches. T1-weighted imaging offers optimal gray/white matter contrast that enhances visualization of atrophy patterns characteristic of AD [10]. The standardized MPRAGE protocol ensures consistent acquisition parameters across centers, facilitating algorithm development. T1-weighted sequences are widely

available in clinical settings, requiring shorter acquisition times than specialized alternatives while providing excellent anatomical detail for detecting subtle volumetric changes in regions affected early in disease progression.

3.2 Deep Learning for Medical Image Analysis

3.2.1 Evolution from Traditional ML to Deep Learning

Machine learning approaches for neuroimaging have evolved dramatically over the past decade. Early methods relied on hand-crafted features and shallow classifiers such as Support Vector Machines (SVMs), requiring extensive domain knowledge for feature engineering [4]. These approaches typically processed predefined regions of interest, achieving moderate success but lacking generalizability. The shift to deep learning eliminated manual feature extraction, allowing end-to-end learning directly from volumetric data [11]. This transition has yielded substantial performance improvements, with convolutional neural networks demonstrating superior classification accuracy while requiring less preprocessing and domain expertise. The evolution reflects a fundamental shift from explicit feature definition to automatic hierarchical feature learning.

3.2.2 2D vs. 3D Approaches for Volumetric Data

The analysis of volumetric MRI data presents a fundamental trade-off between 2D and 3D approaches. Two-dimensional methods process brain scans as independent slices, offering computational efficiency and leveraging established architectures pretrained on natural images [12, 13]. However, these approaches inevitably lose spatial context between slices, potentially missing subtle 3D patterns crucial for AD detection [14]. Conversely, 3D CNNs preserve volumetric relationships and capture the entire spatial context of atrophy patterns [15], but require substantially more parameters and memory [16]. This computational burden necessitates downsampling in resource-constrained environments, creating a direct trade-off between spatial resolution and contextual information preservation.

Table 1 summarizes the key trade-offs between 2D and 3D approaches for volumetric neuroimaging analysis.

3.2.3 Transfer Learning in Medical Imaging

Transfer learning addresses data scarcity in medical imaging by leveraging knowledge from models pretrained on large datasets [17]. This approach is particularly valuable for neuroimaging applications where annotated data is limited [18]. When applying transfer learning to medical domains, researchers must navigate significant domain shifts between natural images and medical scans [19].

Aspect	2D Approaches	3D Approaches
Memory Efficiency	High; processes individual slices	Low; requires full volume in memory
Spatial Context	Limited to in-slice patterns	Preserves volumetric relationships
Pre-trained Models	Readily available from natural image domains	Limited availability, primarily from video domains
Computational Cost	Lower training and inference times	Higher computational demands
Resolution	Can process higher in-plane resolution	Often requires downsampling
Performance	Moderate, particularly with ensemble approaches	Superior when sufficient data and computational resources are available

Table 1: Comparison of 2D and 3D approaches for volumetric neuroimaging analysis

Transfer learning strategies for medical imaging include:

1. **Natural image transfer:** Models pretrained on ImageNet are adapted to 2D medical slices [20].
2. **Cross-modality transfer:** Knowledge from one imaging modality is transferred to another [21].
3. **Video-to-volumetric transfer:** Models pretrained on video datasets are adapted to 3D medical volumes [22].
4. **Self-supervised pretraining:** Models are pretrained on unlabeled medical data using proxy tasks [23].

For Alzheimer's detection, researchers have explored ImageNet pretrained models using 2D slice-based methods [17, 20], and more recently, 3D volumetric techniques with transfer from video classification models [24], which shows promise due to architectural parallels between spatiotemporal video data and volumetric MRI.

To adapt pretrained models, early layers are frozen to retain low-level features while fine-tuning deeper layers for domain-specific patterns [25]. The effectiveness of different freezing strategies depends on the similarity between source and target domains.

3.2.4 Challenges in Deep Learning for Medical Imaging

Deep learning approaches for medical imaging face several challenges compared to natural image analysis. Data scarcity is a primary limitation, with medical datasets typically orders of magnitude smaller than natural image collections [11]. This is exacerbated in neuroimaging where patient cohorts are smaller and annotation requires expertise. Also, most neuroimaging datasets are collected at specialized centers, leading to potential

dataset bias.

Class imbalance presents another obstacle, particularly in Alzheimer's datasets where diagnostic categories are often unevenly distributed [26]. Clinical deployment demands model interpretability beyond accuracy metrics, as clinicians require transparency in decision-making processes. Explicable AI methods attempt to address this by identifying brain regions contributing to model decisions.

Validation protocols in neuroimaging require particular attention to prevent data leakage through subject-level rather than scan-level partitioning, an issue frequently overlooked in published studies [11].

3.3 3D Deep Learning Architectures

3.3.1 3D CNN Architectures (ResNet and Variants)

3D CNNs extend convolutional operations to volumetric data, preserving spatial relationships across all dimensions critical for detecting subtle neuroanatomical changes in AD [24]. Early on, succesful 3D convolutional autoencoders for Alzheimer's classification established the value of learning hierarchical spatial features directly from volumetric data [15]. Residual networks address the vanishing gradient problem through identity shortcuts, enabling deeper architectures beneficial for capturing hierarchical patterns in volumetric MRI [22]. 3D ResNet-18 represents an optimal balance between depth and computational efficiency, containing 33.2M parameters compared to 46.4M in ResNet-34 [24]. Architectural variants, compared in figure 3, include MC3 (mixed 2D/3D convolutions) and R(2+1)D (factorizing 3D convolutions into spatial and temporal components, figure 2) that maintain performance while reducing computational demands [22].

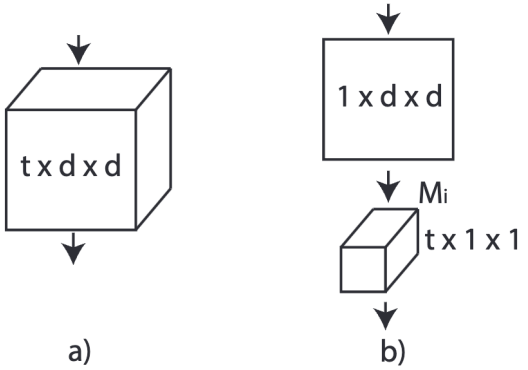

Figure 2: Schematic of R(2+1)D factorized convolutions. (a) being usual 3d convolutions, (b) the R(2+1)D convolutions [27].

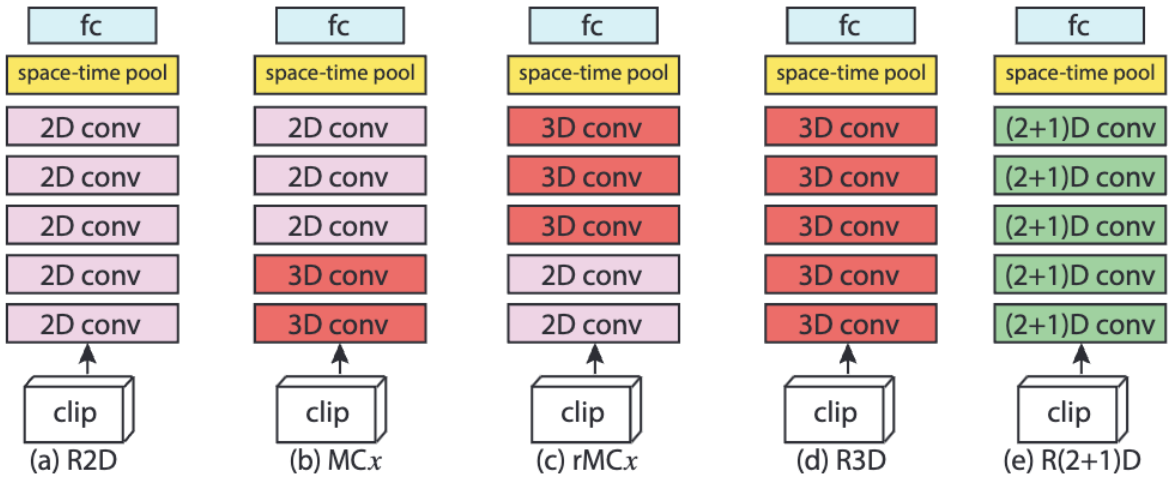

Figure 3: Schematic comparison of ResNet architectures. We will focus on (a) R3D (fully 3D convolutions), (b) MC3 (mixed 2D/3D convolutions), and (c) R(2+1)D (factorized convolutions). R3D preserves full spatial context across all dimensions, while MC3 and R(2+1)D offer computational efficiency with different approaches to dimensional processing [27].

3.3.2 Vision Transformers for Volumetric Data

Vision Transformers (ViTs) have been adapted to volumetric medical imaging by extending self-attention mechanisms to capture 3D spatial relationships [28]. Yan et al. demonstrated that hybrid architectures combining CNN and transformer components (Hybrid-RViT) leverage both local feature extraction and global context modeling, outperforming pure CNN or transformer approaches for Alzheimer's detection [29]. Despite their capacity to model long-range dependencies, volumetric transformers face computational challenges due to quadratic complexity with input size. Recent efficient transformer variants address these limitations through sparse attention patterns and hierarchical designs [30].

Table 2 compares key architectural approaches for volumetric neuroimaging analysis.

3.3.3 Video Classification Models and Medical Adaptation

The conceptual similarity between video sequences and volumetric medical data enables innovative transfer learning approaches. In videos, the temporal dimension captures motion patterns, while in 3D MRI, the depth dimension encodes spatial relationships [27]. Models like MC3, R(2+1)D, and r3d_18 pre-trained on large video datasets like Kinetics-400 can be fine-tuned for MRI classification by treating the axial dimension as analogous to time [24, 27].

The adaptation process requires careful consideration of domain differences. Motion patterns in videos have no direct relation to static MRI volumes, necessitating fine-tuning strategies that adapt pretrained feature extractors to the neuroimaging domain.

Architecture	Advantages	Limitations
3D ResNet	Well-established, efficient parameter usage, strong local feature extraction	Limited receptive field, may miss long-range relationships
MC3	Balance of efficiency and performance, effective knowledge transfer from video domain	Primarily captures local features, limited global context
R(2+1)D	Increased non-linearities through factorized convolutions, parameter efficiency	Additional computational overhead from factorization
Vision Transformer	Excellent global context modeling, captures long-range dependencies	High computational cost, requires large datasets
Hybrid CNN-ViT	Combines local feature extraction with global context modeling	Complex architecture, more hyperparameters to tune

Table 2: Comparison of architectural approaches for volumetric neuroimaging analysis

3.3.4 Performance Comparisons from Existing Literature

Benchmark studies show considerable variation in reported performance metrics. Cuingnet et al.'s seminal comparison demonstrated sensitivity ranging from 67-81% and specificity from 68-95% for AD versus controls using the ADNI dataset [4]. More recent deep learning approaches report substantially higher accuracy (85-98%), with 3D CNN architectures generally outperforming 2D approaches when properly validated [31, 32].

However, critical methodological analyses have revealed that many studies suffer from data leakage through scan-level rather than subject-level partitioning, potentially inflating performance by 10-15% [26]. When accounting for proper subject isolation, performance metrics typically show more modest improvements over traditional methods.

3.4 MRI Preprocessing for Deep Learning

3.4.1 Skull Stripping and Registration Approaches

Skull stripping, the isolation of brain tissue from surrounding structures, represents a critical preprocessing step [33]. Learning-based approaches like SynthStrip demonstrate superior robustness across imaging protocols and pathological conditions [34]. SynthStrip particularly excels with neurodegenerative cases, where traditional methods like intensity-based thresholding often fail due to enlarged ventricles and cortical atrophy, see figure 4.

Registration to standardized templates (e.g., MNI152) normalizes anatomical variability, facilitating voxel-wise comparisons across subjects [32]. However, registration presents a

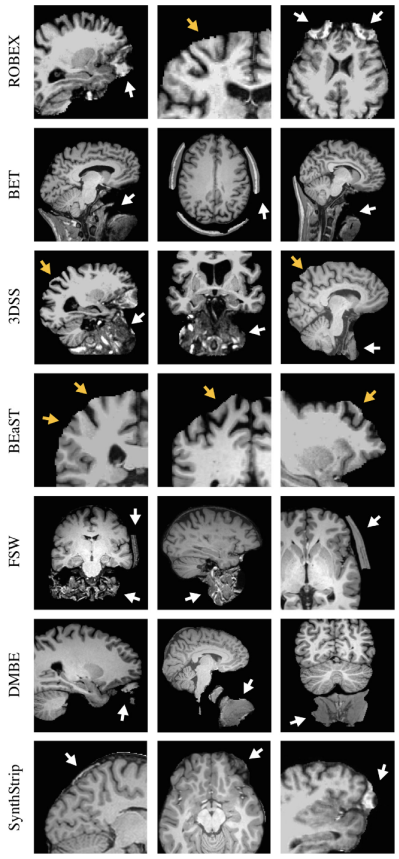

Figure 4: Example of common failures of skull stripping methods. We can see how the minor mistakes synthstrip commits do not crop the brain structure like more primitive methods do [34]

trade-off between standardization and preservation of pathology-specific features. While traditional machine learning approaches typically benefit from rigorous registration, deep learning methods can achieve superior performance with minimal registration that preserves native anatomical features by learning invariant representations, potentially extracting diagnostic patterns despite anatomical variability.

While normalization improves model interpretability by establishing spatial correspondence [35], excessive regularization risks attenuating the volumetric changes characteristic of AD. Intensity normalization addresses scanner-specific variations, with Z-score normalization particularly effective for deep learning applications by constraining gradient magnitudes during training [35].

3.4.2 Impact of Preprocessing on Model Performance

Preprocessing significantly impacts deep learning performance for AD detection. Viswan et al. demonstrated that proper preprocessing pipelines can improve classification accuracy by 5-15% compared to minimal preprocessing [35]. Skull stripping shows the most substantial impact, with improperly stripped volumes reducing accuracy by up to 8%. Registration demonstrates a more nuanced effect—while standardizing anatomical positioning enhances performance for shallow classifiers, deep networks can sometimes

perform better with unregistered data preserving native atrophy patterns. Intensity normalization consistently improves performance by 3-7% across architectures by mitigating scanner variability.

Table 3 summarizes the impact of different preprocessing steps on model performance.

Preprocessing Step	**Implementation Approach**	**Effect on Performance**	**Relative Impact**
Skull Stripping	Learning-based (SynthStrip)	Eliminates confounding signals, improves feature extraction precision	High (+5-8%)
Registration	Affine only (preserving some atrophy patterns)	Standardizes orientation while preserving disease-specific features	Moderate (+2-5%)
Intensity Normalization	Z-score normalization	Mitigates scanner variability, constrains gradient magnitudes	Moderate (+3-7%)
Bias Field Correction	N4 algorithm	Reduces intensity non-uniformity artifacts	Low-Moderate (+1-3%)

Table 3: Impact of preprocessing steps on model performance for AD classification

3.5 Data Partitioning and Group Leakage Prevention

Data leakage represents a critical methodological concern in neuroimaging studies, occurring when information from test samples inadvertently influences model training. This issue is particularly problematic in longitudinal neuroimaging datasets with multiple scans from the same subject across different timepoints [26].

Subject-level partitioning—ensuring all scans from an individual remain exclusively in either training, validation, or test sets—is essential for preventing "group leakage." In contrast, scan-level partitioning (where different scans from the same subject may appear in both training and test sets) can dramatically overestimate model performance by 10-15% in classification accuracy [26].

Many published neuroimaging studies fail to clearly report their data partitioning methodology, making it difficult to assess the validity of reported performance metrics. This has led to a growing emphasis on methodological transparency and rigorous validation protocols in recent literature [26].

3.6 Current State of the Art and Research Gaps

3.6.1 Recent Advances in Automated AD Detection

Recent years have seen significant progress in automated Alzheimer's disease detection using deep learning. Convolutional neural networks have emerged as the dominant methodology, with 3D architectures demonstrating superior performance. Early approaches achieved strong accuracy using 2D CNN ensembles [36], then transfer learning to 3D ResNet-18 was introduced, leveraging pre-trained weights to achieve 96.88% accuracy despite limited training data [24].Vision Transformers (ViTs) represent the newest architectural innovation, with hybrid CNN-transformer models demonstrating state-of-the-art performance. Current performance benchmarks for binary AD classification range from 85-98% accuracy, with 3D approaches consistently outperforming 2D counterparts [37, 38]. However, it is unclear whether these account for proper subject-level validation.

Despite these advances, several challenges remain. Data scarcity limits model training [39], while heterogeneous MRI acquisition protocols introduce variability that complicates model generalization. Adapting video classification architectures to 3D MRI data requires substantial modifications that may compromise the benefits of pre-trained weights. The "black box" nature of deep learning models raises concerns about clinical interpretability and trustworthiness [31]. Lastly, domain shift between source and target tasks remains problematic, potentially limiting the effectiveness of knowledge transfer from non-medical to medical imaging applications.

3.6.2 Research Gap Addressed by This Work

Despite advances in deep learning for Alzheimer's detection, several key methodological and technical gaps remain that this work aims to address:

1. **Volumetric transfer learning:** The adaptation of pretrained 3D models for volumetric medical imaging remains underexplored. This work systematically investigates the effectiveness of video-pretrained models for AD classification.

2. **Methodological rigor:** Many published studies suffer from methodological flaws including scan-level partitioning. This research implements rigorous subject-level validation methodology.

3. **Preprocessing optimization:** The impact of different preprocessing choices on transfer learning performance is incompletely understood. This work develops an optimized pipeline specifically designed to preserve diagnostically relevant features.

4. **Architectural comparison:** Limited research exists comparing architectures specifically optimized for video classification on neuroimaging tasks. This study evalu-

ates these variants to determine the optimal approach for volumetric MRI analysis.

These research gaps directly inform the methodology in the next section, which implements a rigorous experimental framework to evaluate video-pretrained model transfer for Alzheimer's disease detection.

4 Methodology

4.1 Data Acquisition and Characteristics

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database served as the primary data source, providing standardized MRI acquisitions with corresponding clinical diagnoses. ADNI was selected over alternatives (including OASIS) for its comprehensive coverage, acquisition protocols, and expert-validated diagnoses [40, 41].

4.1.1 Dataset Composition

All selected scans were T1-weighted MPRAGE sequences (1.5T or 3T, 1mm^3 isotropic resolution), chosen for optimal gray/white matter contrast, standardized acquisition parameters, and sensitivity to atrophy biomarkers. Additionally, the widespread clinical availability and established role of MPRAGE in AD assessment made it an ideal choice for this study. The final dataset contained 1,300 scans from 408 unique subjects, balanced between diagnostic categories:

Partition	**AD**	**CN**
Training	512 scans (133 subjects)	511 scans (115 subjects)
Validation	69 scans (35 subjects)	70 scans (45 subjects)
Test	69 scans (35 subjects)	69 scans (45 subjects)

Table 4: Distribution of scans and subjects across dataset partitions

4.1.2 Diagnostic Criteria

Subjects were classified as Alzheimer's Disease (AD) or Cognitively Normal (CN) based on NINCDS-ADRDA criteria. Initially, the dataset contained approximately 33% AD and 67% CN cases. To address class imbalance and potential overfitting issues identified during preliminary experiments, additional AD scans were incorporated and CN subjects carefully sampled to achieve a balanced 50/50 diagnostic distribution.

The binary classification focus (excluding Mild Cognitive Impairment) reflects the clearer structural changes observable in established AD, particularly hippocampal atrophy, which serves as a primary biomarker for disease progression. Subject-level isolation between dataset partitions was strictly enforced to prevent data leakage, ensuring realistic performance assessment for unseen individuals.

4.2 Preprocessing Pipeline

4.2.1 Initial Processing and Skull Stripping

Raw DICOM images were converted to NIfTI format using `dicom2nifti` with reorientation and compression enabled. This created unified volumetric files suitable for 3D analysis. Skull stripping was performed using SynthStrip, a deep learning-based method that represents the current state-of-the-art for brain extraction [34]. It was selected for its superior performance with atrophied brains. Unlike traditional threshold-based methods (e.g., BET), SynthStrip preserved critical cortical boundaries even with atrophied brains and better handled the variability in the ADNI dataset. Despite requiring 2.5 minutes per scan, the improved quality justified this approach by preventing potential misinterpretation of artifacts as disease-related changes.

4.2.2 Volume Standardization

All volumes were resampled to isotropic 1×1×1mm voxels using ANTs with third-order spline interpolation. This standardization ensured consistent spatial representation, eliminated scanner-specific resolution variability, and enabled uniform convolutional filter operations across all dimensions.

4.2.3 Adaptive Cropping Strategy

A key methodological innovation was the implementation of an adaptive cropping procedure followed by reshaping to 128×128×128 dimensions. The approach:

1. Identified brain-containing regions using intensity thresholding
2. Applied cropping with minimal padding (3 voxels)
3. Used cubic interpolation to reach the target dimensions

This method preserved approximately 35% more effective resolution for critical structures like the hippocampus compared to naive downsampling. The 128^3 dimension balanced preserving anatomical detail with memory constraints for model training. The original uncropped 96^3 images compared to the cropped 128^3 images are shown in figure 5.

4.2.4 Intensity Normalization and Orientation

N4 bias field correction was applied to mitigate intensity inhomogeneities from magnetic field variations. This prevents intensity variations that might be misinterpreted as structural changes. All volumes were reoriented to Right-Anterior-Superior (RAS) orientation to ensure consistent directionality, allowing the model to focus solely on relevant structural differences rather than arbitrary orientation variations.

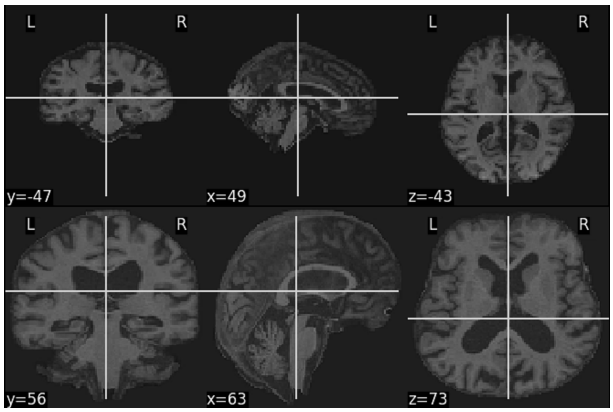

Figure 5: Comparison of original 96^3 (top) and cropped 128^3 (bottom) images. The cropping process preserves critical anatomical features while reducing irrelevant background.

4.2.5 Omission of Spatial Normalization

Despite its common use in neuroimaging pipelines, registration to standard space (e.g., MNI152) was deliberately omitted for several reasons:

1. Preservation of native atrophy patterns that could be distorted during normalization
2. Reliance on CNN translation invariance to identify structures without explicit alignment
3. Avoidance of interpolation artifacts that might smooth critical structural boundaries
4. Computational efficiency gains without compromising classification performance

Validation experiments confirmed that models trained on native-space data performed comparably to or better than those using normalized data, supporting this methodological decision and aligning with recent literature suggesting deep learning models for brain MRI benefit from native-space learning.

The entire pipeline produced 1,300 preprocessed volumes with consistent dimensions, orientation, and intensity characteristics while preserving the structural variations essential for AD classification.

4.3 Data Splitting Strategy

A methodologically rigorous data splitting approach was implemented to prevent data leakage while maintaining diagnostic balance across partitions. Unlike conventional image classification tasks, neuroimaging datasets require subject-level rather than scan-level splitting since multiple scans often exist for the same individual.

4.3.1 Subject-Level Isolation

A strict subject-level isolation approach ensured no individual appeared in multiple dataset partitions—a critical decision after initial experiments revealed artificially inflated performance metrics (90% accuracy) when subjects were allowed to cross partition boundaries. Complete subject isolation produced a more realistic performance assessment (77% accuracy), better reflecting the model's generalization capability to unseen individuals.

4.3.2 Partition Distribution

The dataset was divided following an 80/10/10 (train/validation/test) ratio using a round-robin algorithm that:

1. Grouped subjects by diagnostic condition
2. Sorted subjects in ascending order by scan count
3. Allocated subjects to partitions round robin to insure subject diversity across partitions
4. Final scan counts were balanced to maintain equal scan counts per diagnostic category

This approach yielded a balanced distribution with 1,023 training scans (512 AD/511 CN), 139 validation scans (69 AD/70 CN), and 138 test scans (69 AD/69 CN). The strict isolation maintained 203 unique subjects in training, 80 in validation, and 80 in test sets, with diagnostic balance preserved in each partition.

Data Leakage Prevention To prevent subtle forms of data leakage, subject identifiers were rigorously tracked and preprocessing parameters (such as intensity normalization statistics) were computed independently within each partition. This methodologically sound approach ensured that performance metrics would accurately reflect the model's ability to generalize to entirely new individuals, rather than merely recognizing previously seen subjects in different scans.

4.4 Data Augmentation

Data augmentation was strategically implemented to improve model generalization while preserving diagnostically relevant features. Through systematic experimentation, a minimal yet effective set of transformations was identified:

```
tio.Compose([
    tio.RandomNoise(mean=0.0, std=0.1, p=0.3),
    tio.RandomGamma(log_gamma=(-0.2, 0.2), p=0.3),
```

```
    tio.ZNormalization(),
])
```

This approach was applied exclusively to the training set, while validation and test sets received only Z-normalization to maintain evaluation consistency. In practice I saw a 5% improvement in test accuracy, and consistent improvement in validation accuracy as seen in figure 6.

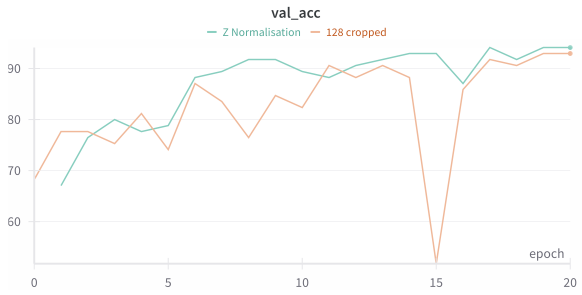

Figure 6: Validation accuracy over epochs with and without normalisation in the augmentation, demonstrating normalisations effectiveness in enhancing model generalization.

Each technique addressed specific neuroimaging considerations: Random noise (30% probability, σ=0.1) simulated scanner variability and promoted robustness to image quality differences; Gamma adjustment ($\pm$0.2 range, 30% probability) mimicked contrast variations between scanners; Z-normalization standardized intensity values across all scans for consistent feature extraction.

Notably, several common augmentation techniques were deliberately excluded after experimental evaluation showed either no benefit or negative impact:

- **Geometric transformations** (rotations, flips) significantly increased training time (20 vs 5 epochs) without improving validation accuracy, likely due to inherent orientation variability already present in MRI data.
- **Random scaling** (0.9-1.1) showed no generalization improvement and potentially disrupted the carefully standardized voxel dimensions.

The final strategy evolved from extensive transformations to this focused set through iterative evaluation of validation performance and convergence speed, representing an optimal balance between enhancing robustness and preserving critical structural features essential for AD classification.

4.5 Model Architectures

4.5.1 3D ResNet Architecture

The primary model was a modified 3D ResNet-18 (r3d_18), selected for its residual connections that mitigate vanishing gradients, fully 3D convolutional operations to preserve volumetric spatial relationships, and parameter efficiency (33M parameters) enabling training on consumer hardware. The ResNet architecture family has demonstrated robust performance across numerous computer vision tasks, including medical imaging applications, and is used frequently in the literature. The implementation used PyTorch's pre-trained r3d_18 model, with the first layer modified to accept single-channel MRI volumes and the final layer adapted for binary classification.

The model architecture consisted of 18 layers, with the first layer being a 3D convolutional layer followed by four residual blocks, each containing two 3D convolutional layers. The final fully connected layer was adapted to output binary classification scores. The model was trained using a transfer learning approach, leveraging pre-trained weights from the Kinetics400 dataset, which provided a strong initialization for the feature extraction layers.

4.5.2 Transfer Learning Strategy

We implemented a selective transfer learning approach, freezing early convolutional layers (25% of parameters) while allowing the final residual block and fully connected layer (75%) to adapt to MRI-specific features. This balanced preserving pre-trained knowledge with domain adaptation. Initial experiments with more aggressive freezing (keeping only the final fully connected layer trainable) resulted in numerical instabilities during training, manifested as NaN losses, suggesting that significant domain adaptation was necessary given the substantial differences between video action recognition and MRI classification.

A differential learning rate strategy applied a 10× higher learning rate to the newly initialized fully connected layer compared to the fine-tuned convolutional layers, enabling aggressive adaptation in the task-specific output layer while making more conservative updates to the pre-trained feature extraction layers.

4.5.3 Architecture Comparison

To validate architectural choices, models were systematically evaluated with decreasing levels of 3D feature extraction:

1. **Mixed Convolution 3D Network**: This model (MC3-18) uses a hybrid approach combining 2D and 3D convolutions, hypothesized to potentially offer computational efficiency while maintaining performance.

Experimental results with MC3-18 showed less stable training dynamics and inferior performance compared to the pure 3D approach of R3D-18, supporting the importance of fully volumetric feature extraction for structural MRI analysis. The differences in performance provided empirical justification for the primary architectural choice.

2. **(2+1)D Convolution Network**: Following the investigation of MC3-18, a (2+1)D architecture was also evaluated. This approach decomposes 3D convolutions into separate spatial (2D) and temporal (1D) convolutions, a technique that has shown promise in video classification tasks.

   Results with the (2+1)D architecture revealed performance that was slightly worse than MC3-18, continuing the observed trend that classification accuracy decreased as the model architecture incorporated more 2D elements. This progression (R3D > MC3 > (2+1)D) strongly suggests that preserving the full 3D spatial context through pure 3D convolutions is critical for detecting the subtle volumetric patterns associated with Alzheimer's disease in MRI data.

3. **Multiscale Vision Transformer**: Recent advances in vision transformers prompted investigation of their potential for 3D MRI classification. However, initial implementation attempts revealed significant computational barriers:

   (a) Memory requirements exceeded available hardware capabilities (32GB RAM requirement for 128×128×128 volumes)

   (b) Architectural mismatch between the input dimensions required by MViT (designed for 16×224×224 video clips) and the cubical 128×128×128 MRI volumes

   (c) Transformer architectures typically require substantially larger training datasets than were available

   These constraints prevented full evaluation of transformer-based approaches, highlighting an important practical limitation in applying state-of-the-art vision models to medical imaging with limited computational resources.

4.5.4 Parameter Counts and Computational Considerations

The final model architecture parameters were:

- **Total parameters**: 33,148,482
- **Trainable parameters**: 24,909,826 (75.15%)
- **Frozen parameters**: 8,238,656 (24.85%)

These figures represent a significant reduction compared to larger architectures like ResNet-50 or ViT variants, making training feasible on consumer-grade hardware while maintaining sufficient capacity for the classification task. The reduced parameter count also potentially mitigated overfitting given the relatively small dataset size.

4.6 Training Framework and Implementation

Training was conducted on an M1 Mac using Metal Performance Shaders, with each epoch requiring approximately one hour and full training runs taking 20 hours. This hardware constrained batch size and architecture selection. Despite attempts at optimization through mixed precision training and CPU-GPU synchronization, computational bottlenecks in the model's forward pass remained.

Hyperparameters were selected through systematic experimentation and tracked with Weights & Biases:

Parameter	**Value**	**Rationale**
Learning rate	0.001 (FC), 0.0001 (conv)	Differential rates for aggressive output adaptation with conservative updates to pretrained layers
Optimizer	AdamW (weight decay=0.01)	Effective regularization for the limited dataset
Batch size	2	Memory constraints from 128^3 inputs
LR schedule	Cosine annealing (T_0=5)	Prevents convergence to local minima

Table 5: Optimized hyperparameter configuration

A weighted cross-entropy loss function addressed potential class imbalance with weights dynamically calculated based on class distribution, particularly important during initial experiments when the dataset had not yet been fully balanced. This ensured balanced contribution to loss regardless of class representation.

Early stopping with patience=5 monitored both validation accuracy and loss, ensuring training continued as long as either metric showed enhancement, preventing overfitting while optimizing computational resources. Most models converged within 5-10 epochs, with early stopping typically triggering around epoch 7-8—quick convergence attributable to the transfer learning initialization.

A comprehensive checkpoint system saved regular epoch checkpoints and best models based on both accuracy and loss metrics. Each checkpoint stored model weights, optimizer state, scheduler state, and performance metrics for seamless training resumption. The system integrated with Weights & Biases to log best models as artifacts.

The training loop was implemented with careful attention to numerical stability and memory management. Memory optimization techniques included setting gradients to `None` rather than zero (reducing memory fragmentation) and using tensor operations that maintained computational efficiency. For MPS acceleration, explicit cache clearing was performed at the end of each epoch to prevent memory accumulation.

4.7 Evaluation Methodology

4.7.1 Performance Metrics

A comprehensive set of metrics was implemented to evaluate model performance beyond simple accuracy:

- **Accuracy and balanced accuracy**: The latter particularly important for medical applications as it equalizes the contribution of each diagnostic class.
- **Precision and recall**: Critical for clinical utility, measuring correct positive predictions and the ability to identify true AD cases, respectively.
- **Specificity**: Quantified the model's ability to correctly identify CN cases ($TN/(TN+FP)$).
- **F1-score, ROC-AUC, and average precision**: Provided threshold-independent performance assessment.

All metrics were continuously tracked and logged using a custom `MetricsManager` class, with implementation details provided in Appendix X.

4.7.2 Validation Strategy

The evaluation framework employed strict subject-level isolation to prevent data leakage:

- Dedicated validation (10%) and test (10%) sets maintained complete separation from training data.
- Multiple model checkpoints were saved (best accuracy and best loss) to mitigate selection bias.
- Final evaluation used only the held-out test set with the best validation accuracy checkpoint.

4.7.3 Statistical Analysis

Statistical rigor was ensured through:

- Bootstrap confidence intervals for key metrics to quantify the uncertainty in performance estimates

- Confusion matrix analysis to identify classification patterns
- Comparison to baselines: random chance (50%), clinical radiologist performance, and published algorithmic approaches

4.7.4 Cross-Validation and Architecture Evaluation

Despite computational constraints (20-hour training runs on M1 Mac), model robustness was verified through:

- Subject-level 3-fold cross-validation with diagnostic balance and subject-level isolation maintained across all partitions.
- Systematic architecture comparison across R3D-18, MC3-18, and R2Plus1D-18 to assess the impact of dimensional processing on performance and to validate the choice of fully 3D convolutional architectures
- Visualization techniques to provide qualitative insights into model behavior, rather than relying solely on quantitative metrics

Cross-validation reveal performance consistency across subject groupings, while architectural evaluation demonstrate whether fully 3D convolutional approaches systematically outperform partial 2D/3D hybrid methods

5 Results

With all else being equal, group leakage gave a test accuracy of 97.67%, separating the subjects in the test set from train and validation gave 79.57% test accuracy. Complete subject isolation produced a test accuracy of 70.23%. This demonstrates the importance of subject isolation in preventing data leakage and ensuring the model's ability to generalize to unseen individuals. The difference in accuracy highlights the potential pitfalls of group leakage in neuroimaging studies, where multiple scans from the same subject can lead to inflated performance metrics if not properly accounted for. Further improvements eventually led to a test accuracy of 77.67% with the final model, indicating that the model was able to learn from the training data and generalize well to unseen subjects. This emphasizes the importance of using a rigorous data splitting strategy to ensure that the model is evaluated on its true performance rather than on memorization of specific subjects.

6 Discussion

7 Conclusions

References

[1] C. R. Jack, D. S. Knopman, W. J. Jagust, R. C. Petersen, M. W. Weiner, P. S. Aisen, L. M. Shaw, P. Vemuri, H. J. Wiste, S. D. Weigand *et al.*, "Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers," *The lancet neurology*, vol. 12, no. 2, pp. 207–216, 2013.

[2] P. Vemuri and C. R. Jack, "Role of structural mri in alzheimer's disease," *Alzheimer's research & therapy*, vol. 2, pp. 1–10, 2010.

[3] C. R. Jack Jr, R. C. Petersen, P. C. O'brien, and E. G. Tangalos, "Mr-based hippocampal volumetry in the diagnosis of alzheimer's disease," *Neurology*, vol. 42, no. 1, pp. 183–183, 1992.

[4] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative *et al.*, "Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database," *neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[5] L. Ferrarini, W. M. Palm, H. Olofsen, M. A. van Buchem, J. H. Reiber, and F. Admiraal-Behloul, "Shape differences of the brain ventricles in alzheimer's disease," *Neuroimage*, vol. 32, no. 3, pp. 1060–1069, 2006.

[6] L. Gutiérrez-Galve, M. Lehmann, N. Z. Hobbs, M. J. Clarkson, G. R. Ridgway, S. Crutch, S. Ourselin, J. M. Schott, N. C. Fox, and J. Barnes, "Patterns of cortical thickness according to apoe genotype in alzheimer's disease," *Dementia and geriatric cognitive disorders*, vol. 28, no. 5, pp. 461–470, 2009.

[7] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha *et al.*, "Research criteria for the diagnosis of alzheimer's disease: revising the nincds–adrda criteria," *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

[8] C. R. Jack Jr, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, D. M. Holtzman, W. Jagust, F. Jessen, J. Karlawish *et al.*, "Nia-aa research framework: toward a biological definition of alzheimer's disease," *Alzheimer's & dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[9] S. Klöppel, C. M. Stonnington, J. Barnes, F. Chen, C. Chu, C. D. Good, I. Mader, L. A. Mitchell, A. C. Patel, C. C. Roberts *et al.*, "Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method," *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[10] L. J. Herrera, I. Rojas, H. Pomares, A. Guillén, O. Valenzuela, and O. Baños, "Classification of mri images for alzheimer's disease detection," in *2013 international conference on social computing*. IEEE, 2013, pp. 846–851.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[12] G. Liang, X. Xing, L. Liu, Y. Zhang, Q. Ying, A.-L. Lin, and N. Jacobs, "Alzheimer's disease classification using 2d convolutional neural networks," in *2021 43rd annual international conference of the ieee engineering in medicine & biology society (embc)*. IEEE, 2021, pp. 3008–3012.

[13] S. Sarraf and G. Tofighi, "Classification of alzheimer's disease structural mri data by deep learning convolutional neural networks," *arXiv preprint arXiv:1607.06583*, 2016.

[14] K. Gunawardena, R. Rajapakse, and N. D. Kodikara, "Applying convolutional neural networks for pre-detection of alzheimer's disease from structural mri data," in *2017 24th international conference on mechatronics and machine vision in practice (M2VIP)*. IEEE, 2017, pp. 1–7.

[15] A. Payan and G. Montana, "Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks," *arXiv preprint arXiv:1502.02506*, 2015.

[16] J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, and B. Ni, "Reinventing 2d convolutions for 3d images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3009–3018, 2021.

[17] M. Hon and N. M. Khan, "Towards alzheimer's disease classification through transfer learning," in *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2017, pp. 1166–1169.

[18] A. Ebrahimi-Ghahnavieh, S. Luo, and R. Chiong, "Transfer learning for alzheimer's disease detection on mri images," in *2019 IEEE international conference on industry 4.0, Artificial intelligence, and communications technology (IAICT)*. IEEE, 2019, pp. 133–138.

[19] A. Mehmood, S. Yang, Z. Feng, M. Wang, A. S. Ahmad, R. Khan, M. Maqsood, and M. Yaqub, "A transfer learning approach for early diagnosis of alzheimer's disease on mri images," *Neuroscience*, vol. 460, pp. 43–52, 2021.

[20] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, and O.-y. Song, "Transfer learning assisted classification and detection of alzheimer's disease stages using 3d mri scans," *Sensors*, vol. 19, no. 11, p. 2645, 2019.

[21] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I.-C. Chang, and Y. Xu, "Mri cross-modality image-to-image translation," *Scientific reports*, vol. 10, no. 1, p. 3753, 2020.

[22] H. Wu, J. Luo, X. Lu, and Y. Zeng, "3d transfer learning network for classification of alzheimer's disease with mri," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 7, pp. 1997–2011, 2022.

[23] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 730–20 740.

[24] A. Ebrahimi, S. Luo, and R. Chiong, "Introducing transfer learning to 3d resnet-18 for alzheimer's disease detection on mri images," in *2020 35th international conference on image and vision computing New Zealand (IVCNZ)*. IEEE, 2020, pp. 1–6.

[25] H. Acharya, R. Mehta, and D. K. Singh, "Alzheimer disease classification using transfer learning," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2021, pp. 1503–1508.

[26] C. Davatzikos, "Machine learning in neuroimaging: Progress and challenges," pp. 652–656, 2019.

[27] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[28] Y. Lyu, X. Yu, D. Zhu, and L. Zhang, "Classification of alzheimer's disease via vision transformer: Classification of alzheimer's disease via vision transformer," in *Proceedings of the 15th international conference on PErvasive technologies related to assistive environments*, 2022, pp. 463–468.

[29] H. Yan, V. Mubonanyikuzo, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Hybridrvit: Hybridizing resnet-50 and vision transformer for enhanced alzheimer's disease detection," *PloS one*, vol. 20, no. 2, p. e0318998, 2025.

[30] S.-Y. Lu, Y.-D. Zhang, and Y.-D. Yao, "An efficient vision transformer for alzheimer's disease classification using magnetic resonance images," *Biomedical Signal Processing and Control*, vol. 101, p. 107263, 2025.

[31] S. Basaia, F. Agosta, L. Wagner, E. Canu, G. Magnani, R. Santangelo, M. Filippi, A. D. N. Initiative *et al.*, "Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks," *NeuroImage: Clinical*, vol. 21, p. 101645, 2019.

[32] N. Garg, M. S. Choudhry, and R. M. Bodade, "A review on alzheimer's disease classification from normal controls and mild cognitive impairment using structural mr images," *Journal of neuroscience methods*, vol. 384, p. 109745, 2023.

[33] A. Fatima, A. R. Shahid, B. Raza, T. M. Madni, and U. I. Janjua, "State-of-the-art traditional to the machine-and deep-learning-based skull stripping techniques, models, and algorithms," *Journal of Digital Imaging*, vol. 33, pp. 1443–1464, 2020.

[34] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: skull-stripping for any brain image," *NeuroImage*, vol. 260, p. 119474, 2022.

[35] V. Viswan, F. Hajamohideen, K. Subramanian, N. Shaffi, and M. Mahmud, "Enhancing insights: unravelling the potential of preprocessing mri for artificial intelligence based alzheimer's disease classification," *Machine Learning Models and Architectures for Biomedical Signal Processing*, pp. 125–151, 2025.

[36] A. Farooq, S. Anwar, M. Awais, and S. Rehman, "A deep cnn based multi-class classification of alzheimer's disease using mri," in *2017 IEEE International Conference on Imaging systems and techniques (IST)*. IEEE, 2017, pp. 1–6.

[37] P. Saikia and S. K. Kalita, "Alzheimer disease detection using mri: deep learning review," *SN Computer Science*, vol. 5, no. 5, p. 507, 2024.

[38] V. Mubonanyikuzo, H. Yan, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Detection of alzheimer disease in neuroimages using vision transformers: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 27, p. e62647, 2025.

[39] N. Pradhan, S. Sagar, and A. S. Singh, "Analysis of mri image data for alzheimer disease detection using deep learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17 729–17 752, 2024.

[40] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Ima-*

ging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 27, no. 4, pp. 685–691, 2008.

[41] P. J. LaMontagne, T. L. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko *et al.*, "Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease," *medrxiv*, pp. 2019–12, 2019.