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: 4th April 2025
Word count: X,XXX

Project supervisor:
Dr Rickson Mesquita

Contents

1 Abstract

2 Introduction

3 Literature Review

3.1 Alzheimer's Disease and Neuroimaging

3.1.1 Pathophysiology with Emphasis on Structural Changes

3.1.2 Hippocampal Atrophy as Primary Biomarker

3.1.3 Additional Neuroimaging Markers

3.1.4 Current Clinical Diagnostic Practices and Limitations

3.1.5 Role of Structural MRI in Diagnosis

3.1.6 Advantages of T1-weighted Imaging for AD Detection

3.2 Deep Learning for Medical Image Analysis

3.2.1 Evolution from Traditional ML to Deep Learning

3.2.2 2D vs. 3D Approaches for Volumetric Data

3.2.3 Transfer Learning in Medical Imaging

3.2.4 Challenges in Deep Learning for Medical Imaging

3.3 3D Deep Learning Architectures

3.3.1 3D CNN Architectures (ResNet and Variants)

3.3.2 Vision Transformers for Volumetric Data

3.3.3 Video Classification Models and Medical Adaptation

3.3.4 Performance Comparisons from Existing Literature

3.4 MRI Preprocessing for Deep Learning

3.4.1 Skull Stripping Methodologies

3.4.2 Registration and Normalization Approaches

3.4.3 Impact of Preprocessing on Model Performance

3.4.4 Current Best Practices

3.4.5 Data Partitioning and Group Leakage Prevention

3.5 Current State of the Art

gnoses. ADNI was selected over alternatives (including OASIS) for its comprehensive coverage, acquisition protocols, and expert-validated diagnoses.

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 1: 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. 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.

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.

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.

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 2: 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

6 Discussion

7 Conclusions

References