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: 1st April 2025
Word count: X,XXX

Project supervisor:
Dr Rickson Mesquita

Contents

1 Abstract

2 Introduction

3 Literature review

4 Methodology

4.1 Data Acquisition and Characteristics

The Alzheimer's Disease Neuroimaging Initiative (ADNI) database served as the primary data source for this study. ADNI represents a comprehensive, longitudinal dataset specifically designed for Alzheimer's disease research, offering rigorously standardized MRI acquisitions with corresponding clinical diagnoses and metadata.

4.1.1 Dataset Selection and Access

After evaluating potential neuroimaging repositories (including OASIS), ADNI was selected for its comprehensive coverage, standardized acquisition protocols, and expert-validated diagnoses. Access was obtained through a formal application process describing the research objectives. The analysis initially utilized data from ADNI-1, later expanding to incorporate volumes from ADNI-2, ADNI-3, and ADNI-4 to increase sample diversity and size.

4.1.2 Image Acquisition Parameters

All selected scans were T1-weighted Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequences, chosen for their optimal gray/white matter contrast which facilitates hippocampal visualization, standardized acquisition parameters across multiple ADNI sites, high signal-to-noise ratio for structural analysis, and sensitivity to hippocampal atrophy, a primary biomarker for AD progression. Additionally, the widespread clinical availability and established role of MPRAGE in AD assessment made it an ideal choice for this study.

The T1w MPRAGE sequences typically featured field strengths of 1.5T or 3T with approximately $1mm^3$ isotropic resolution. The acquisition matrix was approximately $256\times256\times170$, with TR/TE parameters standardized according to ADNI protocol to ensure consistency across imaging sites.

4.1.3 Subject Demographics and Diagnostic Criteria

Subjects were classified into two distinct diagnostic categories: Alzheimer's Disease (AD) and Cognitively Normal (CN). The AD cohort consisted of subjects meeting NINCDS-ADRDA criteria for probable AD, while the CN cohort comprised control subjects without significant cognitive impairment. The original dataset distribution was approximately 33% AD and 67% CN cases. After initial testing revealed potential overfitting issues (discussed later), additional scans were incorporated. During this expansion phase, all available new AD scans were included, with CN subjects carefully sampled to achieve a balanced 50/50 diagnostic distribution in the final dataset to optimize model training.

4.1.4 Data Distribution Analysis

The final dataset contained 1,300 T1w MRI scans from 408 unique subjects:

- AD cohort: 650 scans from 203 subjects
- CN cohort: 650 scans from 205 subjects

Following subject-level splitting (detailed in Data Splitting Strategy section), the distribution across partitions was:

- Training set: 1,023 scans (512 AD, 511 CN) from 248 subjects (133 AD, 115 CN)
- Validation set: 139 scans (69 AD, 70 CN) from 80 subjects (35 AD, 45 CN)
- Test set: 138 scans (69 AD, 69 CN) from 80 subjects (35 AD, 45 CN)

This distribution ensured each partition contained sufficient samples for robust model training and evaluation while maintaining diagnostic balance. The deliberate focus on AD versus CN classification (excluding Mild Cognitive Impairment) reflects the clearer structural changes observable in established AD, particularly the hippocampal atrophy that serves as a primary biomarker for disease progression.

4.2 Preprocessing Pipeline

The preprocessing pipeline was meticulously designed to prepare structural MRI data for optimal deep learning model performance while preserving clinically relevant features. Each stage was selected based on neuroimaging best practices and computational considerations specific to 3D neural network training.

4.2.1 DICOM to NIfTI Conversion

The initial step involved converting the native DICOM format files from ADNI to NIfTI format. This conversion was essential as NIfTI provides a consolidated volumetric rep-

resentation of brain scans, facilitating 3D processing compared to the slice-by-slice arrangement of DICOM files. The conversion preserved header information while creating unified volumetric files using the `dicom2nifti` library with reorientation applied during conversion to ensure consistent initial alignment.

```
dicom2nifti.convert_directory(root, nii_output_dir, compression=True, reorient=True)
```

This compression parameter was enabled to reduce storage requirements without information loss, particularly important given the large dataset size (1,300 scans).

4.2.2 Skull Stripping

Skull stripping was implemented using SynthStrip, a deep learning-based method that represents the current state-of-the-art for brain extraction. The selection of SynthStrip over traditional alternatives like Brain Extraction Tool (BET) was justified by several key advantages. SynthStrip demonstrates superior robustness to variations across diverse acquisition parameters and pathological conditions, which is critical for a heterogeneous dataset spanning multiple ADNI phases. The deep learning foundation of SynthStrip provides more consistent results across subjects compared to threshold-based methods, as more primitive approaches were shown to inaccurately crop atrophied regions, leading to significant information loss. Additionally, SynthStrip better preserves the detailed cortical boundaries that may contain relevant structural information for AD classification. As a synthetic data-trained model, SynthStrip also handles the variability in ADNI data more effectively than traditional algorithms, offering stronger generalization capability.

While SynthStrip required approximately 2.5 minutes per scan on the available hardware, this processing time was justified by the quality of results, as inconsistent skull stripping could introduce confounding artifacts that might be misinterpreted as disease-related changes.

4.2.3 Voxel Standardization

Spatial resolution standardization was performed using ANTs (Advanced Normalization Tools) to resample all volumes to isotropic $1\times1\times1$mm voxels:

```
resampled_img = ants.resample_image(img, (1,1,1), use_voxels=False)
```

This standardization step was crucial for three primary reasons:

1. **Eliminating resolution variability**: Although ADNI enforces acquisition protocols, some variation in voxel dimensions exists across scanners and timepoints.
2. **Isotropic representation**: Consistent cubic voxels ensure that convolutional filters operate uniformly across all three dimensions, preventing directional bias.

3. **Model compatibility**: Standardized resolution simplifies the implementation of 3D convolutional operations and ensures consistent spatial feature extraction.

The resampling was implemented using third-order spline interpolation to maintain structural integrity during resolution adjustment.

4.2.4 Cropping and Reshaping Strategy

A critical preprocessing innovation was an adaptive cropping procedure followed by reshaping to 128×128×128 dimensions. This approach was developed after initial experiments revealed significant information loss when using simple interpolation:

```
# Crop the brain with padding
cropped_img, crop_coords = crop_brain_from_mri(img_data, padding=3)

# Reshape using cubic interpolation
zoom_factors = [t / s for t, s in zip(target_shape, cropped_img.shape)]
final_img = zoom(cropped_img, zoom_factors, order=3)
```

The implemented method:

1. Automatically identifies brain-containing regions using intensity thresholding
2. Crops to these regions with a 3-voxel padding to ensure complete brain coverage
3. Applies cubic interpolation to the cropped volume to reach target dimensions

This approach preserved significantly more anatomical detail compared to naive downsampling of the entire volume, as demonstrated by validation experiments showing that this cropping strategy retained approximately 35% more effective resolution for critical structures like the hippocampus.

The 128×128×128 dimension was selected based on:

- Sufficient resolution to preserve hippocampal and ventricular details
- Memory constraints for model training
- Compatibility with deep network architectures
- Balanced compromise between resolution and computational efficiency

4.2.5 Bias Field Correction and Orientation Standardization

N4 bias field correction was applied to mitigate intensity inhomogeneities resulting from magnetic field variations:

```
bias_corrected = ants.n4_bias_field_correction(input_image)
```

This correction is particularly important for AD classification as it prevents intensity variations that might be misinterpreted as structural changes. Similarly, all volumes were reoriented to Right-Anterior-Superior (RAS) orientation to ensure consistent directionality across the dataset:

```
canonical_img = nib.as_closest_canonical(img)
```

Standardized orientation eliminates the potential confound of different brain orientations influencing the learning process, allowing the model to focus solely on relevant structural differences.

4.2.6 Spatial Normalization

While conventional neuroimaging pipelines often include registration to a standard template space (e.g., MNI152), this step was deliberately omitted for several key reasons:

1. **Preservation of native atrophy patterns**: Spatial normalization can distort or obscure the very atrophic changes that differentiate AD patients from controls, particularly in the hippocampus.
2. **Model capability**: Deep convolutional networks demonstrate inherent translation invariance and can learn to identify relevant structures regardless of precise alignment, making explicit normalization potentially redundant.
3. **Avoiding interpolation artifacts**: The registration process introduces additional interpolation steps that can smooth subtle structural boundaries critical for classification.
4. **Computational efficiency**: Omitting this intensive processing step significantly reduced preprocessing time without compromising classification performance.

Validation experiments confirmed that models trained on native-space data performed comparably to or better than those trained on normalized data, supporting this methodological decision. This approach is aligned with recent literature suggesting that deep learning models for brain MRI classification benefit from learning in subject-native space rather than standardized space.

The comprehensive pipeline ultimately produced a dataset of 1,300 preprocessed volumes with consistent dimensions, orientation, and intensity characteristics while preserving the structural variations essential for AD classification. This carefully crafted preprocessing strategy balances computational constraints with the preservation of clinically relevant features, providing an optimal foundation for the subsequent neural network training.

4.3 Data Splitting Strategy

The data splitting strategy was carefully designed to prevent data leakage while ensuring balanced representation of diagnostic groups across training, validation, and test sets. Unlike conventional image classification tasks, neuroimaging datasets present unique challenges as multiple scans often exist for the same subject across different timepoints, requiring subject-level rather than scan-level splitting.

4.3.1 Subject-Level Isolation

A strict subject-level isolation approach was implemented to ensure no individual subject appeared in multiple dataset partitions. This critical methodological decision was motivated by initial experiments that revealed artificially inflated performance metrics (90% accuracy) when subjects were allowed to appear across partitions. By completely isolating subjects between splits, a more realistic performance assessment (70% accuracy) was achieved, better reflecting the model's true generalization capability to unseen individuals.

4.3.2 Round-Robin Approach for Balanced Distribution

A round-robin selection algorithm was implemented to ensure balanced representation across dataset partitions while maintaining diagnostic class balance. This approach methodically cycled through subjects, allocating them to train, validation, and test sets according to predetermined ratios (80% training, 10% validation, 10% test) while ensuring an equal number of scans from each diagnostic category:

1. Subjects were first grouped by diagnostic condition (AD or CN).
2. Within each condition, subjects are sorted in ascending order by the number of scans that pertain to them.
3. The round-robin algorithm allocated subjects to each partition, test, then validation, the train, until target scan counts were achieved.
4. Final scan counts were balanced to prevent class imbalance (650 scans per diagnostic category).

This approach ensured that even with minimal data there was a large enough subject diversity in the validation and test sets to give a fair evaluation.

4.3.3 Final Distribution Statistics

The final dataset distribution across partitions after implementing the subject-level isolation and round-robin approach was:

- **Alzheimer's Disease (AD) cohort:**
  - Training set: 512 scans from 133 unique subjects
  - Validation set: 69 scans from 35 unique subjects
  - Test set: 69 scans from 35 unique subjects
- **Cognitively Normal (CN) cohort:**
  - Training set: 511 scans from 115 unique subjects
  - Validation set: 70 scans from 45 unique subjects
  - Test set: 69 scans from 45 unique subjects

This distribution ensured approximately 79% of scans were allocated to training, with the remaining 21% evenly divided between validation and test sets, while maintaining diagnostic balance within each partition.

Data Leakage Prevention Special attention was devoted to preventing subtle forms of data leakage that could compromise model evaluation. The subject-level isolation was rigorously enforced through tracking of unique subject identifiers, and all preprocessing parameters (such as intensity normalization statistics) were computed independently within each partition to prevent information bleeding across splits.

This methodologically sound splitting approach provided a robust foundation for model training and evaluation, ensuring that performance metrics would accurately reflect the model's ability to generalize to entirely new subjects rather than merely recognizing previously seen individuals in different scans.

4.4 Data Augmentation

Data augmentation was strategically implemented to improve model generalization by exposing the network to controlled variations while preserving clinically relevant features. The augmentation pipeline evolved through experimental validation to balance diversity enhancement with preservation of diagnostic information.

4.4.1 Augmentation Strategy Development

The augmentation approach underwent several iterations, beginning with a comprehensive set of transformations adapted from general computer vision practices. Through systematic evaluation, the final pipeline was refined to include only those transformations that demonstrably improved generalization without distorting critical diagnostic features:

```
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 minimalist approach was adopted after observing that more aggressive transformations either failed to improve performance or actively degraded it. The augmentation pipeline was applied exclusively to the training set, while validation and test sets received only intensity normalization to maintain evaluation consistency.

4.4.2 Justification for Selected Techniques

Each augmentation technique was selected based on specific neuroimaging considerations:

1. **Random Noise Addition**
    - Simulates natural scanner variability and noise artifacts
    - Promotes robustness to image quality variations across scanning sites
    - Implemented with a moderate noise level to preserve structural integrity
    - 30% probability prevents overreliance on noise-resilient features
2. **Random Gamma Adjustment**
    - Simulates intensity variations common in MRI acquisition
    - Enhances model robustness to contrast differences between scanners
    - Restricted to a narrow range to preserve anatomical relationships
    - Complements the bias field correction applied during preprocessing
3. **Z-Score Normalization** (applied to all volumes)
    - Standardizes intensity values to zero mean and unit variance
    - Critical for consistent feature extraction across scans
    - Mitigates the effect of scanner-specific intensity scales
    - Applied to all datasets (not just training) to ensure consistent input distribution

4.4.3 Augmentation Impact Analysis

Notably, several common augmentation techniques were deliberately excluded after experimental evaluation showed they either provided no benefit or negatively impacted

performance:

1. **Geometric Transformations** (rotations, flips):
    - Initial experiments included rotations ($\pm 90°$) and random flips
    - These transformations significantly increased training time ($\sim$20 epochs vs. $\sim$5 epochs to converge)
    - Provided no measurable improvement in validation accuracy
    - Likely redundant given the inherent orientation variability already present in MRI data
    - May have introduced unrealistic transformations not encountered in clinical settings
2. **Random Scaling**:
    - Initially tested with scale factors of 0.9-1.1
    - Showed no significant improvement in generalization
    - Potentially disrupted the carefully standardized voxel dimensions established during preprocessing

The progression from extensive transformations (using MONAI's comprehensive augmentation library) to a more focused set (using TorchIO's targeted medical imaging augmentations) and finally to the minimal set described above reflected an evidence-based refinement process. This evolution was guided by systematically monitoring validation performance and convergence speed after each modification.

The final augmentation strategy represents an optimal balance between enhancing model robustness and preserving the clinically significant structural features essential for accurate AD classification, particularly the hippocampal atrophy patterns that serve as primary biomarkers.

5 Results

6 Discussion

7 Conclusions

References