# Journal Pre-proof

Adapting to evolving MRI data: A transfer learning approach for Alzheimer's disease prediction

Rosanna Turrisi, Sarthak Pati, Giovanni Pioggia, Gennaro Tartarisco, for the Alzheimer's Disease Neuroimaging Initiative

PII: S1053-8119(25)00016-3
DOI: https://doi.org/10.1016/j.neuroimage.2025.121016
Reference: YNIMG 121016

To appear in: *NeuroImage*

Received date : 25 July 2024
Revised date : 7 January 2025
Accepted date : 9 January 2025

Please cite this article as: R. Turrisi, S. Pati, G. Pioggia et al., Adapting to evolving MRI data: A transfer learning approach for Alzheimer's disease prediction. *NeuroImage* (2025), doi: https://doi.org/10.1016/j.neuroimage.2025.121016.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Revised manuscript clean version Journal Pre-proof Click here to view linked References

# Adapting to evolving MRI data: a Transfer Learning approach for Alzheimer's Disease prediction

Rosanna Turrisi [*a,b], Sarthak Pati[c,d], Giovanni Pioggia[a], Gennaro Tartarisco[a], for the Alzheimer's Disease Neuroimaging Initiative

[a] *Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR) Messina Italy*
[b] *Machine Learning Genoa Center (MaLGa) University of Genoa Genoa Italy*
[c] *Department of Pathology and Laboratory Medicine Indiana University School of Medicine Indianapolis IN USA*
[d] *Medical Research Group MLCommons San Francisco CA USA*

## Abstract

Integrating 3D magnetic resonance imaging (MRI) with machine learning has shown promising results in healthcare, especially in detecting Alzheimer's Disease (AD). However, changes in MRI technologies and acquisition protocols often yield limited data, leading to potential overfitting. This study explores Transfer Learning (TL) approaches to enhance AD diagnosis using a Baseline model consisting of a 3D-Convolutional Neural Network trained on 80 3T MRI scans.

Two scenarios are explored: (A) utilizing historical data to address changes in MRI acquisitions (from 1.5T to 3T MRI), and (B) adapting 2D models pre-trained on ImageNet (ResNet18, ResNet50, ResNet101) for 3D image processing when historical data is unavailable. In both scenarios, two modeling approaches are tested. The *General* Approach involves distinct feature extraction and classification steps, using Radiomic features and TL-based features evaluated with six classifiers. The *Deep* Approach integrates these steps by fine-tuning the pre-trained models for AD diagnosis.

In scenario (A), TL significantly boosts the Baseline's accuracy from 63% to 99%. In scenario (B), Radiomic features better represents 3D MRI than TL-features in the General Approach. Nonetheless, fine-tuning models pre-

*Corresponding author at: Institute for Biomedical Research and Innovation (IRIB), National Research Council of Italy (CNR), Messina, Italy.
Email address: rosanna.turrisi@irib.cnr.it (R. Turrisi).



Journal Pre-proof

Journal Pre-proof

trained on natural images can increase the Baseline's accuracy by up to 12 percentage points, achieving an overall accuracy of 83%.

*Keywords:* Alzheimer's Disease, evolving MRI, Domain Adaptation, Transfer Learning, ACS convolutions

## 1. Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that represents the most prevalent form of dementia, affecting over 30 million people worldwide [1]. It manifests through progressive neurodegeneration, resulting in cognitive and functional decline, impairment of daily activities, and ultimately, mortality [2]. Magnetic Resonance Imaging (MRI) is a fundamental diagnostic tool for the detection of AD, offering detailed insights into structural brain changes associated with the disease. Recently, Machine learning (ML) algorithms have demonstrated a significant potential in leveraging MRI data to enhance the accuracy and efficiency of automatic AD diagnosis [3, 4]. For example, in [5], a Convolutional Neural Network (CNN) trained on 7635 MRI scans reached 99.3% of accuracy in AD classification. Similarly, a CNN model trained on 1409 scans achieved an accuracy of 99.2%[6]. However, the diagnosis accuracy decreases significantly when smaller training datasets are used. For instance, authors in [7] demonstrated that a CNN-based approach achieved 87.5% when trained from scratch on a dataset of 305 images. A variant of a CNN proposed in [8] reached 76% of accuracy in AD diagnosis using an MRI dataset of 515 samples. Authors in [9] employed a combined approach utilizing 3D shearlet transform-based descriptors and ML classifiers on a dataset of 200 MRI volumes, with the best combination achieving 70.5% accuracy in AD classification. This issue, commonly referred to as "data hungriness", is widely recognized in the ML field and has been particularly noted in medical applications [10].
In medicine, the availability of large datasets is often limited due to several factors, such as patient privacy concerns, the high costs associated to data acquisition, and the difficulty of obtaining labeled data. Additionally, variations in MRI scanner modles, acquisition protocols, and image resolutions lead to small and fragmented datasets, further limiting the ability to train high-performing ML models. As hospitals adopt new MRI technologies and refine imaging protocols, researchers face the challenge of managing and integrating both old images and new MRI scans. This often results in a *domain*

Journal Pre-proof

*shift* problem, where the distribution of new MRI data (referred to as *target* data) diverges from the one of old data (referred to as *source* data). This phenomenon has been widely acknowledged and analyzed in several studies [11, 12, 13, 14], where researchers quantified the domain shift due to the heterogeneity among MRI scanners and imaging protocols. These studies consistently reported that applying ML models trained on source MRI data directly to target data can lead to performance degradation, as these models may struggle to generalize effectively to the new data distribution.

To address this issue, commonly referred to as the *Domain Adaptation* (DA) problem, we explore the effectiveness of Transfer Learning (TL) methods. TL [15] is a technique within Deep Learning (DL) [16] that leverages knowledge acquired from one task or dataset to enhance the performance of a related but distinct task or dataset. In ML applications where datasets evolve and new tasks emerge, TL has emerged as a promising solution [17, 18, 19, 20]. By utilizing pre-existing data, TL facilitates faster learning, improved generalization, and adaptability to dynamic environments. The extensive information embedded in existing datasets enables a more efficient and flexible learning approach, promoting enhanced model generalization and performance across a wide range of evolving data scenarios. Specifically, in the context of medical imaging, TL proves particularly useful when dealing with the evolving landscape of MRI data.

Although the DA problem has been extensively discussed in the context of natural images [21, 22, 23, 24, 25], there are only a few studies addressing DA in medical image analysis [26] and specifically in AD. In their work [27], the authors consider MRI scans from the ADNI corpus as source dataset and an external dataset as limited target set. They propose an algorithm based on instance weighting [28] to simultaneously leverage source and target samples, aiming to improve AD diagnosis performance on the target set. Another study, reported in [29], tackles the DA problem using functional MRI data. This study involves a small target dataset collected at a local hospital, with the ADNI dataset serving as the source. The approach focuses on projecting and aligning the source and target datasets into a common subspace, where their differences are minimized. Despite their findings highlighting the need for methods to address the DA problem in AD, the approaches in [27, 29] are limited to vector representations, requiring a feature extraction step. Additionally, these methods assume a similarity between source and target datasets [30]. Furthermore, although samples from the source and

Journal Pre-proof

target datasets are collected by different machines and exhibit different distributions, they share the same resolution.

In our study, we propose a novel TL-based framework for improving AD prediction using evolving MRI data. Our approach has been validated on a target dataset comprising a small collection of three-dimensional T1-weighted MRI scans acquired at 3T, a subset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [31]. We focus our work on two distinct scenarios:

- Scenario (A) deals with differences arising from the evolution of MRI technology between source and target data, including changes in scanner models and resolution with a field strength transition from 1.5T to 3T MRI [11].
- Scenario (B), where historical data is absent, relies on two-dimensional natural images as our source data.

We thoroughly explore TL to address both scenario (A) and (B). In scenario (A), we leverage a 3D CNN pre-trained on 1.5T MRI scans, which we refer to as ADnet or ADnetEx when the classification layer is discarded [32]. In scenario (B), where comparable data or models are lacking, we utilize models pre-trained on ImageNet—specifically ResNet18, ResNet50, and ResNet101 [33]. Before leveraging TL, these models are adapted to 3D image processing using Axial-Coronal-Sagittal (ACS) convolutions [34]. In both scenarios, we explore two modeling approaches. The General Approach involves separate feature extraction and classification steps. We obtain MRI representations by employing pre-trained models as feature extractors or by computing radiomic features. These extracted features then serve as input for a classifier. Specifically, we discriminate between Cognitive Normal (CN) subjects and AD patients using K-nearest neighbors (Knn), Random Forest (RF), and Support Vector Machine (SVM) with linear, polynomial, sigmoid, or Radial Basis Function (RBF) kernels. The Deep Approach, on the other hand, simultaneously performs feature extraction and classification. It consists of fine-tuning pre-trained models directly to optimize the task at hand. Figure 1 summarizes the TL-based experiments conducted in this study.

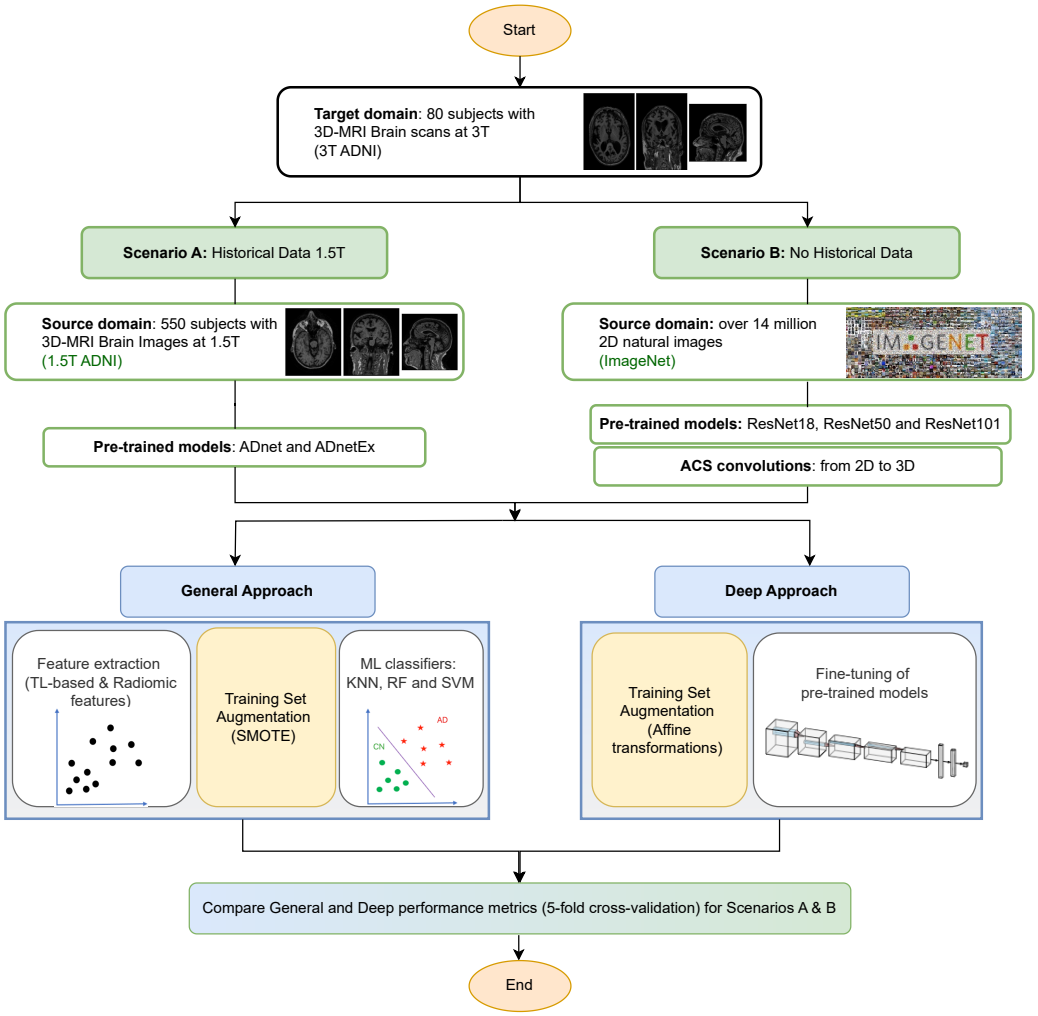


Figure 1: **Overview of the method.** Given the 3T ADNI dataset as target domain, two scenarios are addressed. Scenario (A) employs 1.5 ADNI dataset as source domain, and ADnet and ADnetEx networks as pre-trained models. Scenario (B) overcomes the absence of historical data by leveraging ImageNet as source domain and ResNet-based architectures as pre-trained models. Additionally, ACS convolutions are used to adapt these models to 3D image processing. In both cases, General and Deep Approach are explored and compared.

Journal Pre-proof

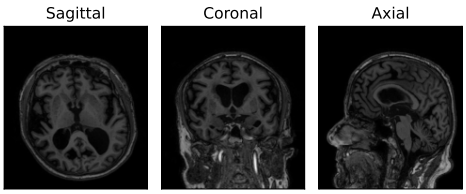


Figure 2: **3T ADNI data.** Axial, coronal and sagittal planes of a 3T MRI scan.

Table 1: **Dataset Demographics.** Number of subjects, averaged age and sex information of 3T and 1.5 ADNI datasets for CN and AD groups.

| **3T** | **CN** | **AD** |
|---|---|---|
| Subjects | 47 | 33 |
| Age | $75.1 \pm 3.9$ | $74.0 \pm 8.1$ |
| Sex (M/F) | 18/29 | 11/22 |

| **1.5T** | **CN** | **AD** |
|---|---|---|
| Subjects | 307 | 243 |
| Age | $75.2 \pm 7.6$ | $75.9 \pm 5.0$ |
| Sex (M/F) | 159/148 | 130/113 |

## 2. Materials and Methods

### *2.1. Dataset*

The Alzheimer's Disease Neuroimaging Initiative (ADNI) [31] comprises heterogeneous datasets collected during different temporal phases (ADNI1, ADNI/GO, ADNI2, and ADNI3), each characterized by changing MRI acquisition protocols. In this study, we focus on MRI scans from the ADNI1 phase present images acquired by various scanner manufacturers, including Phillips System, GE, and Siemens, at different field strengths. All exams have been pre-processed by ADNI experts to harmonize data across these different scanners. Depending on the acquisition system, specific MRI corrections have been applied: image geometry distortion (using Gradwarp), image intensity non-uniformity (via B1 calibration and N3), and image scaling for gradient drift using phantom data [31]. However, despite these efforts, MRI scans from ADNI1 still exhibit significant variability due to differences in scanner models, resolution changes, and varying magnetic field strengths. In [11], the authors quantified the domain shift caused by scanner models and magnetic field strength variability within ADNI1. They specifically compare five Siemens scanner models: Trio, Allegra, Avanto, Sonata, and Symphony. Trio and Allegra scanners operate at a magnetic field strength of 3T, whereas the other three scanners operate at 1.5T. Their findings indicate that MRI exams conducted at 1.5T with the Avanto, Sonata, and Symphony models share similarities in their feature space, suggesting comparable imaging characteristics. In contrast, 1.5 MRI scans and data from the Trio and Allegra

Journal Pre-proof

scanners at 3T exhibit a notable domain shift, with a mean discrepancy distance of 7. To address the domain adaptation challenge posed by changes in MRI protocols, we consider MRI scans acquired by different scanner models at 3T and 1.5T as the target and source datasets, respectively.

Specifically, the target dataset comprises 3T MRI scans from the ADNI1 phase collected during the baseline exam (Fig. 2). As reported in Table 1, this dataset includes 80 3D T1-weighted MRI exams of the brain from 47 Cognitive Normal (CN) subjects (F=29, age=75.1±3.9) and 33 patients with AD (F=22, age=74.0±8.1). We denote this dataset as "3T ADNI dataset".

As the source dataset in scenario (A), we utilized 1.5T MRI scans from the ADNI1 phase, collected during screening. This dataset consists of 550 3D-MRI exams of the brain at 1.5T (Table 1), including 307 CN subjects (F=148, age=75.2±7.6) and 243 patients with AD (F=130, age=75.9±5.0). We refer to this source dataset as the "1.5T ADNI dataset."

In scenario (B), where historical data is unavailable, we utilize the ImageNet dataset as our source data. ImageNet is an extensive collection of over 14 million high-resolution, labeled 2D natural images, covering diverse categories. It encompasses more than 20,000 categories, each containing several hundred images. The annotations in ImageNet adhere to the WordNet hierarchy. Since 2010, the ImageNet project has hosted the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which serves as a benchmark for image classification and object detection.

### *2.2. Pre-trained models*

In this section, we introduce the pre-trained models used in this work, considering scenarios (A) and (B).

*Scenario (A).* A source dataset similar to the target one or a pre-trained model trained on it is available. This is the case for ADnet and ADnetEx models. ADnet is a pre-trained 3D-CNN originally proposed in [32]. Its architecture comprises 8 convolutional layers followed by a feedforward layer, and it was trained on the 1.5T MRI ADNI dataset to distinguish between healthy subjects and patients with AD. In ADnetEx, only the convolutional layers (i.e., the feature extractor) are used as a pre-trained model, and a classification layer is added with randomly initialized weights.

Journal Pre-proof

*Scenario (B).* In situations where we lack access to similar data, we turn to models pre-trained on 2D images, specifically those trained on ImageNet. We considered three state-of-the-art models with increasing numbers of trainable parameters: ResNet18, ResNet50, and ResNet101 [33]. It is important to note that these architectures are based on 2D convolutions, while our target set consists of 3D images. To address this limitation, we adopt ACS convolutions, as proposed in [34]. This approach involves splitting the 2D kernels into three parts (Axial-Coronal-Sagittal) by channels and convolving each part with the 3D input, which has been shown to work well for various healthcare imaging tasks [35, 36].

More precisely, let $X$ be a 3D input of dimension $C_i \times D \times H \times W$, where $C_i$ is the number of channels, and let $W$ be the pre-trained 2D kernel of dimension $C_o \times C_i \times K \times K$, where $C_o$ denotes the number of output channels. The ACS convolution splits $W$ by the output channel into

$$\begin{aligned} W_a &\in \mathbf{R}^{C_o^a \times C_i \times 1 \times K \times K} \\ W_c &\in \mathbf{R}^{C_o^c \times C_i \times K \times 1 \times K} \\ W_s &\in \mathbf{R}^{C_o^s \times C_i \times K \times K \times 1}, \end{aligned}$$

with $C_o^a = C_o^c = C_o^s = C_o/3$. To compute the ACS view-based 3D features, $X$ is separately convolved with $W_a$, $W_c$, $W_s$. Finally, the three outputs are concatenated to obtain the output features.

Table 2 provides information about the input data and the model size of all pre-trained models.

### *2.3. Feature extraction and data processing*

#### *2.3.1. Radiomic features*

Radiomic features have been employed by the scientific community to quantify various imaging properties, particularly in the healthcare domain [37, 38, 39, 40, 41, 42, 43, 44]. We begin the process of extracting Radiomic features by segmenting the Volume of Interest (VOI) to exclude non-brain tissues and the skull. Addressing challenges posed by close spatial proximity and similar intensities of non-brain tissues, we employed a deep learning-based image segmentation approach using the open-source Python

Journal Pre-proof

Table 2: Pre-trained models utilized in this study, providing details on the pre-training dataset (source dataset), input data type, and the number of trainable parameters.

| Scenario | Source | Model | Input | N. of Trainable Parameters |
|---|---|---|---|---|
| A | ADNI 1.5T | ADnet | 3D | 220k |
| | | ADnetEx | 3D | 220k |
| B | ImageNet | ResNet18 | 2D | 11.7 M |
| | | ResNet50 | 2D | 25.6 M |
| | | ResNet101 | 2D | 44.5 M |

tool DeepBrain [45]. DeepBrain relies on a pre-trained CNN architecture based on U-Net for efficient brain extraction [46]. Once we obtained the VOI segmentation, radiomic feature extraction was performed using PyRadiomics, an open-source Python package designed for computing radiomic signatures in medical imaging data [47]. In total, we extracted 107 radiomic features, including 18 features from First Order Statistics, 14 shape-based features, 24 Grey Level Co-occurrence Matrix features, 14 Gray Level Dependence Matrix features, 16 Grey Level Run Length Matrix features, 16 Gray Level Size Zone Matrix features, and 5 Neighboring Gray Tone Difference Matrix features. These radiomic features were exclusively used in the General Approach.

*2.3.2. TL-based features*

In the General Approach, we extracted MRI representations by utilizing pre-trained models as feature extractors, while discarding the classification layer. Specifically, we used ADnet for scenario (A) and ResNet18, ResNet50, and ResNet101 after applying ACS convolutions for scenario (B).

*2.3.3. Data augmentation*

Given that the target dataset contains only 80 samples, we employed data augmentation techniques to expand its size.
In the General Approach, we first extracted the features from the MRI scans and then augmented the dataset. In this case, as we were working with vector data, we utilized the Synthetic Minority Oversampling Technique (SMOTE) [48]. SMOTE is an oversampling method that generates synthetic examples of the minority class within the convex hull of all minority class instances in the training set. Specifically, we randomly selected a sample (S1) from the

Journal Pre-proof

Table 3: General Approach: optimal hyperparameters found via grid search for each classifier and image representation.

| | **Knn** | **RF** | **SVM (L)** | **SVM (P)** | | **SVM (RBF)** | | **SVM (S)** |
|---|---|---|---|---|---|---|---|---|
| **Features** | K | N | C | C | d | C | $\gamma$ | C |
| Radiomics | 3 | 500 | 0.1 | 0.0001 | 2 | 10 | 0.01 | 10 |
| ADnet | 3 | 500 | 0.01 | 10 | 3 | 10 | 0.01 | 10 |
| ADnet+R. | 5 | 100 | 1.0 | 10 | 2 | 10 | 0.01 | 1.0 |
| ResNet18 | 3 | 500 | 0.001 | 0.0001 | 2 | 0.0001 | 0.001 | 0.0001 |
| ResNet50 | 5 | 500 | 0.001 | 10 | 3 | 0.0001 | 0.001 | 0.0001 |
| ResNet101 | 5 | 100 | 0.0001 | 10 | 2 | 0.0001 | 0.001 | 0.0001 |
| ResNet18+R. | 7 | 200 | 0.0001 | 0.0001 | 2 | 0.0001 | 0.001 | 0.0001 |
| ResNet50+R. | 3 | 200 | 0.001 | 0.001 | 2 | 0.0001 | 0.001 | 0.0001 |
| ResNet101+R. | 7 | 100 | 0.0001 | 10 | 2 | 0.0001 | 0.001 | 0.0001 |

minority class, identified k=5 nearest neighbors, and then randomly chose one of them (S2). A synthetic example was created along the line connecting S1 and S2 at a randomly selected point. It is important to note that since the target set is not highly unbalanced, this approach only marginally increases the dataset size.

In the Deep Approach, we applied the data augmentation strategy to the raw images. Following the findings in [32], we applied three different affine transformations (rotation, zoom, and shift) twice, resulting in six images generated from each original MRI scan. In both approaches, data augmentation is performed only on the training set, leaving testing sets at the original sample size.

### *2.4. Experimental details*

#### *2.4.1. General Approach*

We utilized the Radiomic and TL-based features either independently or combined, as input for a classifier, to address the binary classification task of CN vs. AD. Specifically, we employed K-nearest neighbors (Knn), Random Forest (RF), and Support Vector Machine (SVM) classifiers [49] with four different kernels: linear (L), Gaussian (RBF), polynomial (P) and sigmoid (S).

For the ML classifiers, we performed grid search to identify the best hyperparameters. In particular, we employed stratified 5-fold cross-validation, shuffling the dataset with a random state of 42 and setting the test size to 20%. Specifically:

Journal Pre-proof

| Scenario | Model | Pre-training data | Fine-tuning | Training data |
|---|---|---|---|---|
| Baseline | 1.5T_ADnet | – | No | 1.5T ADNI |
| | 3T_ADnet | – | No | 3T ADNI |
| A | ADnet | 1.5T ADNI | Yes | 3T ADNI |
| | ADnetEx | 1.5T ADNI | Yes | 3T ADNI |
| B | ResNet18 | ImageNet | Yes | 3T ADNI |
| | ResNet50 | ImageNet | Yes | 3T ADNI |
| | ResNet101 | ImageNet | Yes | 3T ADNI |

Table 4: Deep Approach: scheme of the training pipeline.

- For Knn, we explored the number of nearest neighbors k varying among 3, 5, and 7.
- In the Random Forest classifier, we searched for the best number of trees N among 5, 100, 200, 500.
- For SVM classifiers, we tuned the regularization parameter C across values of 0.0001, 0.001, 0.01, 0.1, 1, and 10. Additionally, for SVM with the Gaussian kernel, we determined the best kernel coefficient ($\gamma$) from options including 0.001, 0.01, 0.1, 1, and 10. For the polynomial kernel, we investigated polynomial degrees (d) equal to 2 and 3.

Table 3 provides details on the discovered best hyperparameters for each ML method and MRI representation.

*2.4.2. Deep Approach*

This approach involves leveraging a pre-trained model to enhance and expedite the learning task of discriminating between CN and AD subjects in the target dataset. All models, except ADnet, have their last layer removed, and a classification layer is added on top. Additionally, we conducted a preliminary analysis to compare full fine-tuning with freezing the first layer. For all architectures and Scenarios, we found that freezing the first layer yields the best performance. Therefore, in the presented experiments, the first layer of all models was kept frozen. These results are consistent with the findings reported by [50] on large benchmark dataset, which also highlighted that additional layers beyond the first does not significantly impact model performance on large benchmark datasets. Based on these insights, and given

Journal Pre-proof

the substantial domain shift between ImageNet and 3T ADNI, we retained this TL strategy in Scenario B. For Scenario A, however, freezing additional layers could be more appropriate, given the greater similarity between the source and the target datasets. To explore this, we compared model performance in terms of accuracy and required number of epochs for increasing numbers of fine-tuned convolutional layers, ranging from 1 (fine-tuning the last layer) to 7 (fine-tuning all but the first layer).
For comparison, we also trained ADnet with randomly initialized weights on the target dataset without using TL. This model, referred to as "3T_ADnet", serves as a baseline. We chose to perform this training only on the ADnet architecture since it has the smallest number of trainable parameters, and the target dataset is very small. Furthermore, for Scenario (A), we quantified the classification performance of ADnet on the target dataset without any fine-tuning. We refer to this model as "1.5T_ADnet".
All networks were trained by minimizing the cross-entropy loss using the Adam optimizer, with an initial learning rate of 0.0005. An $\ell_2$ penalty of 0.1 was added to prevent overfitting. The batch size was set to 30. Table 4 summarizes all the adopted models.

#### *2.4.3. Performance metrics and cross-validation*

To ensure accurate model assessment, each experiment was set up using stratified-K-fold cross-validation with k=5. At each split, the dataset was divided into a training subset (n∼65) and a testing subset (n∼15). In the General Approach, the training set was augmented using SMOTE, and then both training and testing data underwent normalization. In the Deep Approach, affine transformations were applied to the training set, augmenting it up to 585 samples. Since the target dataset does not have imbalanced classes, we adopted accuracy as the main metric to measure the models' performance. Additionally, the best models were also evaluated in terms of F1-score, precision, recall, Area Under the Curve (AUC), Area Under the Precision-Recall Curve (AUPRC), and confusion matrix.

## 3. Results

### *3.1. Scenario (A)*

To motivate and demonstrate the need for transfer learning (TL), we first quantified the domain shift between the 1.5T and 3T ADNI datasets. For this analysis, we employed the Maximum Mean Discrepancy (MMD) with a

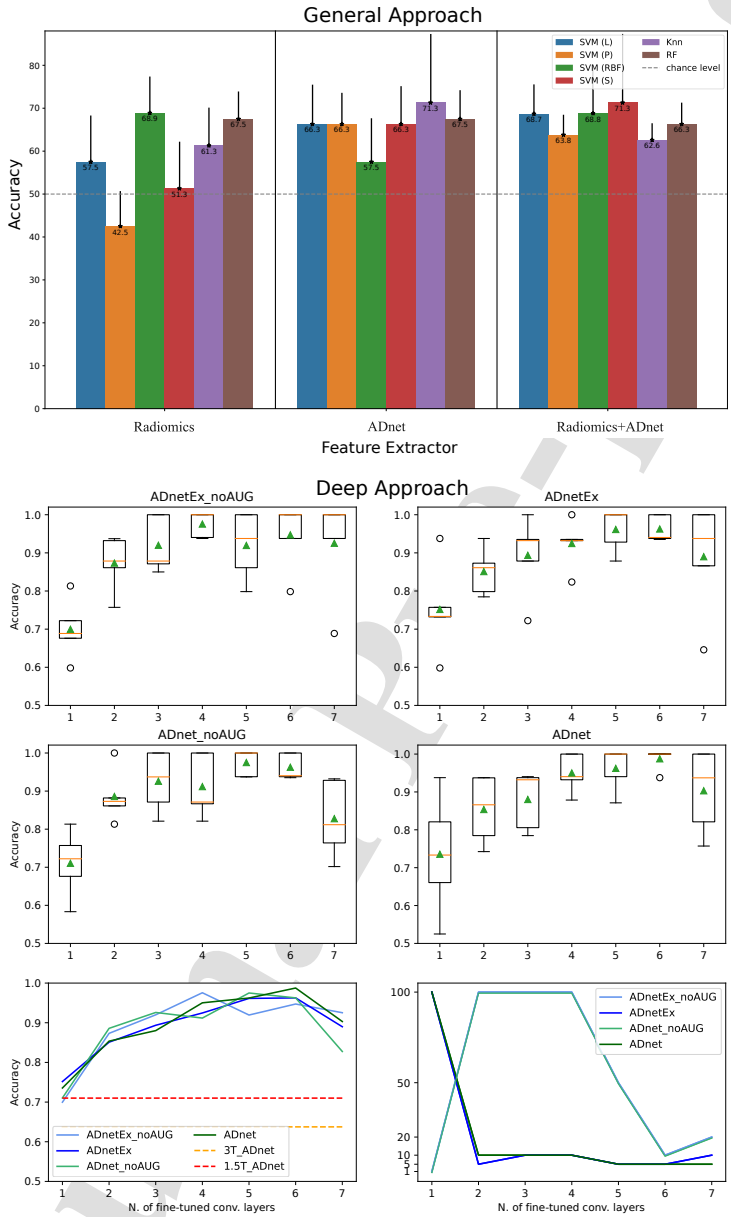


Figure 3: Scenario A. (Top) General approach: accuracy for different input features and classifiers. (Bottom) Deep approach: comparison of fine-tuned models for different number of fine-tuned layers (from 1 to 7).

| | Baseline | Scenario (A) | | | |
|---|---|---|---|---|---|
| **Model** | 3T_ADnet | ADnet | ADnet_noAUG | ADnetEx | ADnetEx_noAUG |
| **Best Epoch** | 100 | 5 | 10 | 5 | 10 |
| **Accuracy** | 62.5 | 98.8 | 96.4 | 96.4 | 95.3 |
| **F1-score** | 51.6 | 98.5 | 95.4 | 95.4 | 93.5 |
| **Precision** | 55.2 | 1.0 | 96.9 | 96.9 | 1.0 |
| **Recall** | 48.5 | 97.0 | 93.9 | 93.9 | 87.9 |
| **AUC** | 58.3 | 98.7 | 99.2 | 97.7 | 99.2 |
| **AUPRC** | 46.5 | 98.8 | 99.0 | 98.3 | 99.1 |

Table 5: Complete evaluation of all models.

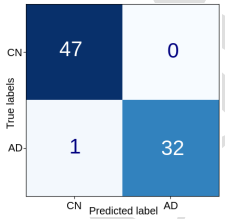


Figure 4: ADnet's confusion matrix.

Gaussian kernel, the Fréchet Inception Distance (FID), and a classifier-based approach. In the classifier-based method, a binary classifier was trained to distinguish between samples from the two datasets. The results were as follows: MMD=0.4, FID=3.3, and classification accuracy=1.0. These findings indicate a moderate domain shift between the source and target domains in Scenario A.

The General Approach evaluated Radiomic features, ADnet-based features, and the combination of them as MRI representations using six classifiers, with a focus on diagnosis accuracy. Our findings, presented in the top panel of Figure 3 , reveal that ADnet-based features consistently outperformed Radiomic features, except for SVM using an RBF kernel. The highest accuracy (71.3±8.5%) was achieved using ADnet-extracted features with Knn. Notably, the same accuracy but reduced standard deviation (71.3±6.4%) was obtained when combining Radiomic and ADnet-based features, indicating that Radiomic features may improve model stability.

In the Deep Approach, we thoroughly investigated the potential of TL by fine-tuning the pre-trained models ADnet and ADnetEx. We also report the performance of these models fine-tuned without any data augmentation strategy, referred to as ADnet_noAUG and ADnetEx_noAUG, respectively. Figure 3 (bottom panel — upper part) displays the box plot accuracy over 5 folds, at increasing number of fine-tuned layers. Plots at the bottom shows the best average accuracy over the folds (on the left) and the number of required epochs (on the right), for ADnet (dark green), ADnetEx (dark blue), ADnet_noAUG (light green) and ADnetEx_noAUG (light blue). To compare TL-based models to the baseline, we reported the accuracy of 3T_ADnet (orange), trained on the target dataset without TL, and the generalization ability of 1.5T_ADnet (red). Since these models do not go under any fine-

Journal Pre-proof

tuning and, therefore, are independent of the number of fine-tuned layers, we have displayed it as a constant values. The results revealed a concave relationship between the number of fine-tuned layers and accuracy, with optimal performance observed when fine-tuning between 4 and 6 layers. For these configurations, the model exhibited similar performance in terms of mean and standard deviation accuracy across k-folds. ADnet and ADnetEX reach their peak accuracy after only a few epochs, whereas a higher number of epochs are necessary for fine-tuning the same models without augmentation. Notably, fine-tuning 6 layers requires the fewest training epochs for all models. Under this configuration, ADnet, ADnetEx, ADnet_noAUG, and ADnetEx_noAUG achieved accuracies of 98.8±2.3%, 96.4±2.9%, 96.4±2.9%, and 95.3±6.8%. Comparing TL-based models to the baselines, we find that 1.5T_ADnet reaches 71% accuracy, while 3T_ADnet achieves up to 62.5±18.4% accuracy. The standard deviation for 3T_ADnet is notably high, highlighting the model's lack of robustness when trained exclusively on the target dataset. Therefore, the best TL-based model outperform 1.5T_ADnet and 3T_ADnet baselines of 27.8 and 36.3 points. A complete evaluation, in terms of required training epochs, accuracy, F1-score, Precision, Recall, AUC and AUPRC is reported in Table 5 for all models. Furthermore, Fig. 4 displays the confusion matrix for the best model (ADnet).

By isolating each component of the TL process, we can observe the following: (i) using a pre-trained feature extractor without fine-tuning outperformed the baseline by 10 points (from 71.3% to 61.3%); (ii) fine-tuning the feature extractor resulted in a significant accuracy gain of 2.4 points (from 71.3% to 95.3%); (iii) pre-initializing the classification layer further enhanced accuracy by 1.1 point (from 95.3% to 96.4%); and (iv) applying data augmentation reduced the number of training epochs (from 10 to 5) while boosting accuracy by 2.4 points (from 96.4% to 98.8%).

### *3.2. Scenario (B)*

This challenging scenario aims to improve the performance of AD diagnosis on a small set of 3D MRI scans in the absence of historical data, using 2D natural images as leverage. Results from the General Approach indicate that, without fine-tuning, models pre-trained on ImageNet struggle to provide an accurate representation of MRI images, achieving a maximum accuracy of 61.3±10.0%. This discrepancy is reasonable given the significant domain shift between the source and target datasets. In this context, the

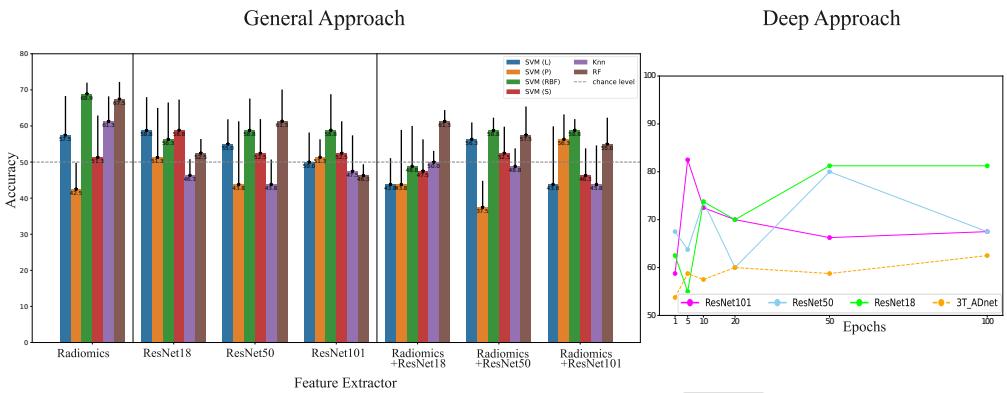


Figure 5: Scenario B. (left) General approach: accuracy for different input features and classifiers. (right) Deep approach: comparison of the baseline models and fine-tuned models for different cases across multiple epochs (1, 5, 10, 20, 50 and 100)

best accuracy (68.9±6.8%) is achieved relying on Radiomic features solely with SVM (RBF) classifier yields an accuracy. Notably, the highest accuracies achieved by classifiers combining Radiomic features with ResNet18-, ResNet50-, and ResNet101-based features were 61.3±4.7%, 58.8±3.1%, and 58.8±3.1%, respectively. This indicates that incorporating radiomic information resulted in accuracy variations of +2.5, -2.5, and 0 points for the three feature types, respectively. Thus, in Scenario B, only the ResNet18-based classifier benefitted from the additional features. One possible explanation is that ResNet18 features have a dimensionality of 512, while ResNet50 and ResNet101 features are 2048-dimensional vectors. Given our dataset's limited size of 80 samples, adding features to already high-dimensional vectors may impair classification performance. Detailed performance metrics for all models are presented in the left panel of Figure 5.

Surprisingly, the Deep Approach reveals that TL remains essential even when dealing with distinct source and target datasets. We report the accuracy averaged over 5 folds, obtained by converting 2D pre-trained models into 3D ACS models and fine-tuning them. Results for fine-tuning strategies without augmentation are not reported due to significantly poorer performance, either underperforming the baseline or achieving only marginally higher accuracy. Figure 5 (right panel) illustrates the performance of the 3T_ADnet (highlighted in orange), which serves as the baseline and reaches up to 71%. Among the ACS-based architectures, fine-tuning ResNet18,

Journal Pre-proof

| | Baseline | Scenario (B) | | |
|---|---|---|---|---|
| **Model** | 3T_ADnet | ResNet18 | ResNet50 | ResNet101 |
| **Best Epoch** | 100 | 50 | 50 | 5 |
| **Accuracy** | 62.5 | 81.3 | 80.0 | 82.5 |
| **F1-score** | 51.6 | 76.27 | 73.3 | 77.4 |
| **Precision** | 55.2 | 80.0 | 81.5 | 82.3 |
| **Recall** | 48.5 | 72.7 | 66.7 | 72.7 |
| **AUC** | 58.3 | 84.9 | 78.3 | 86.9 |
| **AUPRC** | 46.5 | 72.4 | 64.8 | 83.0 |

Table 6: Complete evaluation of all models.

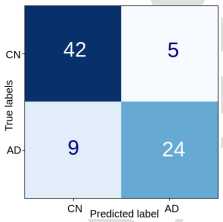


Figure 6: ResNet101's confusion matrix.

ResNet50 and ResNet101 models yields accuracies of 81.9±4.6%, 80.0±4.9%, and 82.5±7.7%, respectively. Hence, ResNet101 emerges as the most effective network for AD classification, although results remain comparable across all models. Importantly, each model attains at least 80%, considerably improving upon the baseline up to 12 points– an indication of the approach's effectiveness.A complete evaluation of fine-tuned models is reported in Table 6, whereas in Figure 6 we display the confusion matrix of ResNet101 model, which correspond to the best fine-tuned model in Scenario (B).

By isolating each component of the TL process, we can observe that (i) utilizing a pre-trained feature extractor (ResNet50) without fine-tuning reduced the baseline performance by 0.9 points (from 62.2% to 61.3%); (ii) fine-tuning the feature extractor significantly improved accuracy, boosting performance by 18.7 points (from 61.3% to 80.0%); and (iii) employing a larger network (ResNet101) further enhanced accuracy by 2.5 points (from 80.0% to 82.5%).

## 4. Discussion

The findings of this study underscore the need for TL when addressing changes in MRI data acquisition, especially in cases with limited target data. Specifically, the proposed Baseline model (3T_ADnet), consisting of a 3D-CNN trained solely on the target data, achieved a diagnosis accuracy of 63%. In contrast, TL-based approaches achieved up to 99% and 83% accuracy in scenarios (A) and (B), respectively. This highlights the critical role of TL in clinical settings, where data availability is often restricted. Our investigation demonstrates that leveraging analogous historical data, when feasible, and utilizing pre-trained 2D models in the absence of such data can

Journal Pre-proof

yield significant benefits in AD diagnosis.

In scenario (A), results based on MMD, FID and a classifier-based approach, proved the presence of a domain shift between 1.5T and 3T MRI scans. Furthermore, the baseline model 1.5T_ADnet, trained solely on 1.5T MRI, achieved only 71% accuracy in classifying 3T MRI. Since in [32], authors reported an accuracy of 82% on the 1.5T ADNI testing set, 1.5T_ADnet shows an 11-point decrease in accuracy when generalizing to 3T MRI. This highlights the domain shift inherent in MRI scans acquired at 1.5T and 3T and underscores the limits of generalized diagnosis models in addressing MRI acquisition changes. One potential strategy to tackle this issue could involve enhancing the generalization ability of 1.5T_ADnet. However, highly accurate generalized models typically require extensive datasets –which are scarce in the medical field– and computationally intensive training, which raises environmental sustainability concerns [51, 52]. In contrast, our approach addresses the domain shift by efficiently adapting the generalized model to the target data through fine-tuning over just a few epochs. This enables the creation of a highly accurate, hospital-specific diagnostic model tailored to new acquisition protocols. Furthermore, it is crucial to notice that training a hospital-specific diagnostic model from scratch on the target dataset yields poor performance when working with limited datasets.

Notably, even the generalization accuracy of 1.5T_ADnet surpasses the prediction accuracy achieved by 3T_ADnet. While training on the target domain typically –and intuitively– outperforms generalizing from the source domain, this advantage diminishes significantly when the target domain is very limited in size. Similar trends have been reported across various areas of medical machine learning (see for instance [53]). In our case, the target dataset contains only 80 samples, which leads to poor generalization. Adapting models pre-trained on historical data to new MRI acquisitions reached up to 99% accuracy, allowing us to i) overcome the overfitting problems when using only the target dataset, and ii) improve to the generalization ability of pre-trained models to enhance the AD diagnosis on new MRI collections. We compared two pre-trained 3D-CNN models, ADnet and ADnetEx, which differ in their classification layer: ADnet's last layer is pre-trained on 1.5T ADNI, whereas in ADnetEx's last layer weights are randomly initialized. Their performances were similar, with an accuracy difference of 2.5 and 1.6 points with and without data augmentation. Additionally, training these models without data augmentation led to a slight decrease in performance

Journal Pre-proof

(1.1 points for ADnetEx and 2.4 points for ADnet), even considering the limited size of the target dataset. Data augmentation accelerated the training phase, reaching the learning peak within 5 epochs. This finding could be useful when balancing the memory complexity and the runtime. Comparison between General and Deep Approach clearly showed the superiority of the latter. Nonetheless, leveraging models pre-trained on 1.5T scans as feature extractors provides better MRI representations compared to Radiomic Features. The best combination of ML classifier and TL-based features achieved a classification accuracy of 71.3%, whereas the best AD diagnosis performance based on Radiomic Features achieved 68.9%.

In Scenario (B), we tackled a more complex challenge by leveraging 2D models pre-trained on ImageNet. While this approach is not entirely novel in the context of AD, some previous studies [54, 55, 56, 57] have employed TL based on 2D pre-trained models to enhance AD diagnosis. However, these studies often overlook small target datasets, limiting their experiments to benchmark datasets that do not always reflect real clinical practice. Furthermore, their approaches are confined to 2D images. For instance, some methods require users to select specific slides from 3D MRI data (as seen in [55]), while others use all slices but lose the inherent 3D information. To address this limitation, we propose adapting 2D pre-trained models to 3D images via ACS convolutions, allowing us to utilize pre-trained weights from models designed for 2D datasets while performing inherently 3D operations. This strategy has also been successfully implemented in brain tumor segmentation [35], where authors demonstrated its superiority over 3D models with random weight initialization. Our results revealed that Radiomic features outperformed using TL-features from non-fine-tuned models. However, by applying ACS convolutions and fine-tuning the converted models, the Deep Approach achieved a notable improvement. It boosted the accuracy of the Baseline model 3T_ADnet by up to 12 points, reaching 82.5%. It is important to highlight that models pre-trained on ImageNet typically consists of millions of parameters. While large architectures are commonly used for transfer learning, this approach may not be optimal in the medical domain, given the limited datasets and the associated risk overfitting. To address this concern, we evaluated architectures of increasing size, specifically ResNet18 (12M parameters), ResNet50 (26M parameters), and ResNet101 (45M parameters). Interestingly, our results demonstrated that model size had a minimal impact on final performance, yielding an average accuracy of 81.5% with a low

Journal Pre-proof

standard deviation of 1.0 across architectures. These findings suggest that even in the absence of directly analogous medical data, leveraging pre-trained models from different domains can provide substantial benefits.

The comparison between the General and Deep Approaches further emphasizes the importance of model adaptation and fine-tuning. Specifically, the Deep Approach significantly improved diagnosis accuracy in both scenarios, outperforming the General Approach by 39% in case (A) and 35% in case (B). Nevertheless, results from the General Approach can be particularly useful when labels are unavailable, and the goal is to generate a good representation of the input images. In this context, the left panels of Figure 3 and 5 provide guidelines for choosing the best MRI representation, suggesting ADnet-based features for case (A) and Radiomic features for case (B).

A fair comparison of our results with state-of-the-art studies is challenging due to the absence of a benchmark for evaluation within the ADNI dataset. Furthermore, o the best of our knowledge, this is the first work proposing this type of domain adaptation in the context of AD diagnosis. Overall, our study highlights the efficacy of TL in improving AD prediction models when faced with limited and varied data. The success of TL in both scenarios underscores its potential as a valuable tool in clinical practice, where data constraints and technological advancements continually pose challenges. Most AD studies based on TL primarily focus on using models pre-trained on natural images, neglecting the potential of leveraging historical similar data. Nonetheless, we discuss findings from the most related work. In [58], a dataset of 111 T1 MRI scans from ADNI corpus is used for the AD diagnosis task. The authors employ TL techniques based on 2D models VoxCNN and ResNet, achieving an accuracy of 79% and 80%, respectively. Although their target dataset size is slightly larger than ours, our approach improves their accuracy by 13 points in scenario (A) and 3 points in scenario (B), highlighting the effectiveness of ACS convolutions and the adaptation of 2D models to 3D images. TL based on 2D models is also proposed in [7], where the authors use ResNet101, Xception, and InceptionV3 architectures pre-trained on ImageNet. These models achieved accuracies of 91%, 85% and 94%, respectively. Although these results are comparable to ours, their target dataset included Mental Cognitive Impaired (MCI) subjects in the AD class to increase the dataset size, resulting in a total number of 305 samples–

Journal Pre-proof

almost four times the size of our target dataset. This emphasizes the impact of the dataset size on the final performance, especially when fine-tuning large networks like ResNet101. Comparing results from scenario (A) and (B) reveals that, when dealing with a limited target size, using similar data and smaller models provides significantly better performance– about 10 points of accuracy– compared to state-of-the-art models pre-trained on ImageNet.

While this study has yielded promising results, it is important to acknowledge certain limitations that warrant caution and consideration for future work. A key constraint in our approach was the limited size of the target dataset, consisting of only 80 samples. However, this was an intentional choice made to replicate real-world scenarios in the medical domain, where data availability is often constrained. This small dataset does present challenges for model generalization and increases the risk of overfitting, particularly when fine-tuning complex deep learning models. Furthermore, the limited sample size precluded the implementation of a robust early stopping mechanism, potentially allowing overtraining. Splitting the dataset further for validation purposes would have compounded the issue, resulting in smaller and less representative training sets. This challenge is not unique to TL but also affects training the 3T_ADnet baseline model. Despite this issue, fine-tuning consistently outperforms the baseline models regardless of the number of training epochs. Empirical results indicate that the optimal number of epochs is relatively low (approximately 5-10) in case (A) and around 50 in case (B). Future work will address this by implementing early stopping methods based on local gradient statistics [59]. In Scenario (B), adapting models pre-trained on a natural image dataset, to 3D medical imaging via ACS convolutions proved effective but was inherently limited by the significant domain gap between the source and target datasets. Although models pre-trained on ImageNet are largely available and easily accessible, it may be advantageous to pre-train from scratch a model on available 2D images datasets, such as MedMNIST [60], which may present a smaller domain shift. Furthermore, this study focused exclusively on binary classification between AD and CN subjects, without addressing cases of mild cognitive impairment (MCI). Future experiments will aim to extend this work to include differential diagnosis tasks, which are expected to introduce new challenges due to the more subtle distinctions between categories.

Journal Pre-proof

## 5. Conclusion

This research addresses challenges posed by evolving MRI technologies and limited data, highlighting the efficacy of Transfer Learning for Alzheimer's Disease diagnosis. We consider a realistic scenario with a small target dataset of 80 MRI scans, showing that training a 3D-CNN model from scratch results in only 63% accuracy.

First, we adopt a General Approach, separately performing feature extraction and classification, which can serves as a guideline for selecting the best MRI representation when labels are unavailable. Our findings suggest using models pre-trained on historical data for feature extraction when available, and Radiomic features otherwise.

Next, we use a Deep Approach based on fine-tuning for AD diagnosis. In scenario (A), we address the challenge of changes in MRI field strengths (from 1.5 to 3 Tesla). The domain shift issue is evident, as a 3D-CNN model trained solely on historical data achieves only 71% accuracy on the target dataset. In this context, TL boosts model accuracy to 99%. Scenario (B) tackles the absence of historical data by leveraging 2D models pre-trained on natural images. To preserve 3D information, we adapt the 2D models to 3D image processing using ACS convolutions. Fine-tuned models achieved an overall accuracy of 83%.

While our primary focus lies in AD diagnosis, the principles explored in this work can be readily extended to other healthcare applications, as the medical field typically grapples with limited data availability and evolving data. Within this context, we demonstrated the promise of TL in healthcare, showcasing its ability to overcome the MRI acquisition changes that naturally occur in hospitals.

## Data and code availability

The ADNI data that support the findings of this study are openly available on request at https://adni.loni.usc.edu/.

All the experiments were conducted using Python version 3.8 and PyTorch 1.12.1, running on a Tesla K40c GPU. Python codes necessary to reproduce the experiments are available at https://github.com/rturrisige/TL4ADdiagnosis.

Journal Pre-proof

Journal Pre-proof

## Ethics statement

Ethical approval for ADNI dataset was granted by the review boards of the involved institutions, and all participants provided written informed consent as per the enrollment protocols.

## CRediT authorship contribution statement

**Rosanna Turrisi:** Writing — original draft, Formal analysis, Methodology, Investigation, Conceptualization. **Sarthak Pati:** Writing – review & editing, Conceptualization, Methodology. **Giovanni Pioggia:** Writing – review & editing. **Gennaro Tartarisco:** Writing – review & editing, Funding acquisition, Validation.

## Declaration of competing interest

The authors declare that they have no competing interests.

## Funding

This work was supported by the Italian project ADAPT-MI (Ministero delle Imprese e del Made in Italy — MIMIT, Prog. n. F/310019/01/X56)

## Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: ADNI Acknowledgment List.

## References

[1] Y.-T. Wu, A. S. Beiser, M. M. Breteler, L. Fratiglioni, C. Helmer, H. C. Hendrie, H. Honda, M. A. Ikram, K. M. Langa, A. Lobo, et al., The changing prevalence and incidence of dementia over time—current evidence, Nature Reviews Neurology 13 (6) (2017) 327–339.

Journal Pre-proof

[2] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, et al., Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria, Alzheimer's & Dementia 12 (3) (2016) 292–323.

[3] Q. Zuo, H. Wu, C. P. Chen, B. Lei, S. Wang, Prior-guided adversarial learning with hypergraph for predicting abnormal connections in alzheimer's disease, IEEE Transactions on Cybernetics (2024).

[4] Y. Zong, Q. Zuo, M. K.-P. Ng, B. Lei, S. Wang, A new brain network construction paradigm for brain disorder via diffusion-based graph contrastive learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[5] A. W. Salehi, P. Baglat, B. B. Sharma, G. Gupta, A. Upadhya, A cnn model: Earlier diagnosis and classification of alzheimer disease using mri, in: 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 156–161. doi:10.1109/ICOSEC49089.2020.9215402.

[6] 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 21 (2019) 101645.

[7] H. Ghaffari, H. Tavakoli, G. Pirzad Jahromi, Deep transfer learning–based fully automated detection and classification of alzheimer's disease on brain mri, The British journal of radiology 95 (1136) (2022) 20211253.

[8] U. Senanayake, A. Sowmya, L. Dawes, Deep fusion pipeline for mild cognitive impairment diagnosis, in: 2018 IEEE 15th international symposium on biomedical imaging (isbi 2018), IEEE, 2018, pp. 1394–1997.

[9] S. Alinsaif, J. Lang, 3d shearlet-based descriptors combined with deep features for the classification of alzheimer's disease based on mri data, Computers in Biology and Medicine 138 (2021) 104879. doi:https://doi.org/10.1016/j.compbiomed.2021.104879.
URL https://www.sciencedirect.com/science/article/pii/S0010482521006739

Journal Pre-proof

[10] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, H. Kurdi, Impact of dataset size on classification performance: an empirical evaluation in the medical domain, Applied Sciences 11 (2) (2021) 796.

[11] R. Kushol, A. H. Wilman, S. Kalra, Y.-H. Yang, Dsmri: domain shift analyzer for multi-center mri datasets, Diagnostics 13 (18) (2023) 2947.

[12] M. Dadar, S. Duchesne, C. Group, et al., Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data, NeuroImage 217 (2020) 116928.

[13] H. Lee, K. Nakamura, S. Narayanan, R. A. Brown, D. L. Arnold, A. D. N. Initiative, et al., Estimating and accounting for the effect of mri scanner changes on longitudinal whole-brain volume change measurements, Neuroimage 184 (2019) 555–565.

[14] J. L. Panman, Y. Y. To, E. L. van der Ende, J. M. Poos, L. C. Jiskoot, L. H. Meeter, E. G. Dopper, M. J. Bouts, M. J. van Osch, S. A. Rombouts, et al., Bias introduced by multiple head coils in mri research: an 8 channel and 32 channel coil comparison, Frontiers in neuroscience 13 (2019) 729.

[15] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242–264.

[16] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[17] G. Pinto, Z. Wang, A. Roy, T. Hong, A. Capozzoli, Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives, Advances in Applied Energy 5 (2022) 100084.

[18] M. Talo, U. B. Baloglu, Ö. Yıldırım, U. R. Acharya, Application of deep transfer learning for automated brain abnormality classification using mr images, Cognitive Systems Research 54 (2019) 176–188.

[19] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, J. Pei, Transfer learning for drug discovery, Journal of Medicinal Chemistry 63 (16) (2020) 8683–8694.

Journal Pre-proof

[20] D. Cook, K. D. Feuz, N. C. Krishnan, Transfer learning for activity recognition: A survey, Knowledge and information systems 36 (2013) 537–556.

[21] A. Farahani, S. Voghoei, K. Rasheed, H. R. Arabnia, A brief review of domain adaptation, Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020 (2021) 877–894.

[22] N. Courty, R. Flamary, A. Habrard, A. Rakotomamonjy, Joint distribution optimal transportation for domain adaptation, Advances in neural information processing systems 30 (2017).

[23] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, K. Kim, Image to image translation for domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4500–4509.

[24] R. Turrisi, R. Flamary, A. Rakotomamonjy, M. Pontil, Multi-source domain adaptation via weighted joint distributions optimal transport, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 1970–1980.

[25] Z. Cao, L. Ma, M. Long, J. Wang, Partial adversarial domain adaptation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 135–150.

[26] H. Guan, M. Liu, Domain adaptation for medical image analysis: a survey, IEEE Transactions on Biomedical Engineering 69 (3) (2021) 1173–1185.

[27] C. Wachinger, M. Reuter, A. D. N. Initiative, et al., Domain adaptation for alzheimer's disease diagnostics, Neuroimage 139 (2016) 470–479.

[28] S. Bickel, M. Brückner, T. Scheffer, Discriminative learning for differing training and test distributions, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 81–88.

[29] W. Li, Y. Zhao, X. Chen, Y. Xiao, Y. Qin, Detecting alzheimer's disease on small dataset: a knowledge transfer perspective, IEEE journal of biomedical and health informatics 23 (3) (2018) 1234–1242.

Journal Pre-proof

[30] S. B. David, T. Lu, T. Luu, D. Pál, Impossibility theorems for domain adaptation, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 129–136.

[31] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, L. Beckett, The alzheimer's disease neuroimaging initiative, Neuroimaging Clinics of North America 15 (4) (2005) 869.

[32] R. Turrisi, A. Verri, A. Barla, Deep learning-based alzheimer's disease detection: reproducibility and the effect of modeling choices, Frontiers in Computational Neuroscience 18 (2024) 1360095.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[34] J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, B. Ni, Reinventing 2d convolutions for 3d images, IEEE Journal of Biomedical and Health Informatics 25 (8) (2021) 3009–3018. doi:10.1109/JBHI.2021.3049452.

[35] B. Baheti, S. Pati, B. Menze, S. Bakas, Leveraging 2d deep learning imagenet-trained models for native 3d medical image analysis, in: International MICCAI Brainlesion Workshop, Springer, 2022, pp. 68–79.

[36] S. Pati, S. P. Thakur, İ. E. Hamamcı, U. Baid, B. Baheti, M. Bhalerao, O. Güley, S. Mouchtaris, D. Lang, S. Thermos, et al., Gandlf: the generally nuanced deep learning framework for scalable end-to-end clinical workflows, Communications Engineering 2 (1) (2023) 23.

[37] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, et al., The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology 295 (2) (2020) 328–338.

[38] C. Davatzikos, S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, E. Cohen, et al., Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diag-

Journal Pre-proof

nostics and predictive modeling of clinical outcome, Journal of medical imaging 5 (1) (2018) 011018–011018.

[39] S. Pati, A. Singh, S. Rathore, A. Gastounioti, M. Bergman, P. Ngo, S. M. Ha, D. Bounias, J. Minock, G. Murphy, et al., The cancer imaging phenomics toolkit (captk): technical overview, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5, Springer, 2020, pp. 380–394.

[40] S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, G. Shukla, J. D. Rudie, N. F. Santamaría, A. F. Kazerooni, S. Pati, et al., The university of pennsylvania glioblastoma (upenn-gbm) cohort: advanced mri, clinical, genomics, & radiomics, Scientific data 9 (1) (2022) 453.

[41] P. Whybra, A. Zwanenburg, V. Andrearczyk, R. Schaer, A. P. Apte, A. Ayotte, B. Baheti, S. Bakas, A. Bettinelli, R. Boellaard, et al., The image biomarker standardization initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights, Radiology 310 (2) (2024) e231319.

[42] R. Chitalia, S. Pati, M. Bhalerao, S. P. Thakur, N. Jahani, V. Belenky, E. S. McDonald, J. Gibbs, D. C. Newitt, N. M. Hylton, et al., Expert tumor annotations and radiomics for locally advanced breast cancer in dce-mri for acrin 6657/i-spy1, Scientific data 9 (1) (2022) 440.

[43] S. Pati, R. Verma, H. Akbari, M. Bilello, V. B. Hill, C. Sako, R. Correa, N. Beig, L. Venet, S. Thakur, et al., Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the ivy glioblastoma atlas project (ivy gap) dataset, Medical physics 47 (12) (2020) 6039–6052.

[44] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, E. Neri, A deep look into radiomics, La radiologia medica 126 (10) (2021) 1296–1311.

[45] A. Kori, M. Soni, B. Pranjal, M. Khened, V. Alex, G. Krishnamurthi, Ensemble of fully convolutional neural network for brain tumor segmentation from magnetic resonance images, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 485–496.

Journal Pre-proof

[46] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[47] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H. J. Aerts, Computational radiomics system to decode the radiographic phenotype, Cancer research 77 (21) (2017) e104–e107.

[48] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[49] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their applications 13 (4) (1998) 18–28.

[50] P. D. Alfano, V. P. Pastore, L. Rosasco, F. Odone, Fine-tuning or top-tuning? transfer learning with pretrained features and fast kernel methods, arXiv preprint arXiv:2209.07932 (2022).

[51] K. Hao, Training a single ai model can emit as much carbon as five cars in their lifetimes (2019), URL https://www. technologyreview. com/2019/06/06/239031 (2019).

[52] C.-J. Wu, B. Acun, R. Raghavendra, K. Hazelwood, Beyond efficiency: Scaling ai sustainably, IEEE Micro (2024).

[53] R. Turrisi, L. Badino, Interpretable dysarthric speaker adaptation based on optimal-transport, in: Interspeech 2022, 2022, pp. 26–30. doi:10.21437/Interspeech.2022-36.

[54] T. M. Ghazal, G. Issa, Alzheimer disease detection empowered with transfer learning, Computers, Materials & Continua 70 (3) (2022) 5005–5019.

[55] N. M. Khan, N. Abraham, M. Hon, Transfer learning with intelligent training data selection for prediction of alzheimer's disease, IEEE Access 7 (2019) 72726–72735.

Journal Pre-proof

[56] H. Acharya, R. Mehta, 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.

[57] A. Ashraf, S. Naz, S. H. Shirazi, I. Razzak, M. Parsad, Deep transfer learning for alzheimer neurological disorder detection, Multimedia Tools and Applications (2021) 1–26.

[58] S. Korolev, A. Safiullin, M. Belyaev, Y. Dodonova, Residual and plain convolutional neural networks for 3d brain mri classification, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 835–838. doi:10.1109/ISBI.2017.7950647.

[59] M. Mahsereci, L. Balles, C. Lassner, P. Hennig, Early stopping without a validation set, arXiv preprint arXiv:1703.09580 (2017).

[60] J. Yang, R. Shi, B. Ni, Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis, in: IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195.

Journal Pre-proof

Journal Pre-proof

# Data and code availability statement

**Adapting to evolving MRI data: a Transfer Learning approach for Alzheimer's Disease prediction**

Rosanna Turrisi[1], Sarthak Pati, Giovanni Pioggia, Gennaro Tartarisco, for the Alzheimer's Disease Neuroimaging Initiative

The ADNI data that support the findings of this study are openly available on request at https://adni.loni.usc.edu.

All the experiments were conducted using Python version 3.8 and PyTorch 1.12.1, running on a Tesla K40c GPU. Python codes necessary to reproduce the experiments are available at https://github.com/rturrisige/TL4ADdiagnosis/.

[1] Corresponding author at: Institute for Biomedical Research and Innovation (IRIB), National Research Council (CNR), Messina, Italy.
Email address: rosanna.turrisi@irib.cnr.it (R. Turrisi).

*Preprint submitted to NeuroImage*

Journal Pre-proof

Journal Pre-proof

# Title and authors information

## Adapting to evolving MRI data: a Transfer Learning approach for Alzheimer's Disease prediction

Rosanna Turrisi[1a,b], Sarthak Pati[c,d], Giovanni Pioggia[a], Gennaro Tartarisco[a], for the Alzheimer's Disease Neuroimaging Initiative

[a]Institute for Biomedical Research and Innovation (IRIB) National Research Council of Italy (CNR) Messina Italy

[b]Machine Learning Genoa Center (MaLGa) University of Genoa Genoa Italy

[c]Department of Pathology and Laboratory Medicine Indiana University School of Medicine Indianapolis IN USA

[d]Medical Research Group MLCommons San Francisco CA USA

[1] Corresponding author at: Institute for Biomedical Research and Innovation (IRIB), National Research Council (CNR), Messina, Italy.
Email address: rosanna.turrisi@irib.cnr.it (R. Turrisi).

*Preprint submitted to NeuroImage*

Journal Pre-proof

Journal Pre-proof

# Highlights

**Adapting to evolving MRI data: a Transfer Learning approach for Alzheimer's Disease prediction**

Rosanna Turrisi[1], Sarthak Pati, Giovanni Pioggia, Gennaro Tartarisco, for the Alzheimer's Disease Neuroimaging Initiative

- Changes MRI technologies and protocols cause domain shifts in collected MRI data
- Transfer Learning enhance Alzheimer's Disease diagnosis with evolving MRI data
- Transfer Learning-based models boost Alzheimer's Disease prediction accuracy from 63% to 99% (Scenario A)
- Fine-tuned 2D models achieve 83% diagnosis accuracy for 3D MRI data processing (Scenario B)

[1] Corresponding author at: Institute for Biomedical Research and Innovation (IRIB), National Research Council (CNR), Messina, Italy.
Email address: rosanna.turrisi@irib.cnr.it (R. Turrisi).

*Preprint submitted to NeuroImage*

Journal Pre-proof

Journal Pre-proof

# Declaration of competing statement

**Adapting to evolving MRI data: a Transfer Learning approach for Alzheimer's Disease prediction**

Rosanna Turrisi[1], Sarthak Pati, Giovanni Pioggia, Gennaro Tartarisco, for the Alzheimer's Disease Neuroimaging Initiative

The authors declare that they have no competing interests.

Journal Pre-proof

1 Corresponding author at: Institute for Biomedical Research and Innovation (IRIB), National Research Council (CNR), Messina, Italy.
Email address: rosanna.turrisi@irib.cnr.it (R. Turrisi).

*Preprint submitted to NeuroImage*