Domain Adaptation of MRI Scanners

by

Rafsanjany Kushol

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science
University of Alberta

Abstract

Deep learning (DL) has become a leading subset of machine learning (ML) and has been successfully employed in diverse areas, ranging from natural language processing to medical image analysis. In medical imaging, researchers have progressively turned towards multi-center neuroimaging studies to address complex questions in neuroscience, leveraging larger sample sizes and aiming to enhance the accuracy of DL models. However, challenges arise due to variations in imaging characteristics across centers, often attributed to differences in MRI scanners. This phenomenon, known as domain shift, leads to inconsistent performance of DL models when applied to unknown test data. Domain adaptation (DA) methods aim to bridge this domain gap by aligning data across different domains. Unfortunately, the lack of suitable tools for domain shift analysis hinders the development and validation of DA techniques. Moreover, existing solutions often process entire datasets without accounting for source/target domain heterogeneity. Furthermore, the impact of various MRI scanners on different disease classification tasks remains largely unexplored.

Motivated by the aforementioned challenges and limitations of existing literature, I first propose a novel framework called DSMRI (Domain Shift analyzer for MRI) to comprehensively assess the extent of domain shift within MRI datasets. This framework provides key insights into domain shift factors by integrating knowledge from diverse domains, including spatial, frequency, wavelet, and texture analysis. Secondly, I introduce another unsupervised framework called DeepDSMRI, which analyzes domain shift in MRI data using various deep models pre-trained on the ImageNet dataset. DeepDSMRI demonstrates its efficacy in determining domain shift

not only in structural MRI (*e.g.*, T1-weighted, T2-weighted, and FLAIR) but also in advanced MRI modalities such as diffusion-weighted imaging (DWI) and functional MRI (fMRI). To the best of my knowledge, this is the first work to analyze and quantify domain shift in multi-modal MRI using DL without requiring additional training on MRI data.

Thirdly, I investigate the impact of scanner vendor variability on various disease classification tasks using multiple DL models. My analysis reveals a significant decline in classification accuracy when DL models are tested with data from different scanner manufacturers. To address the challenging task of amyotrophic lateral sclerosis (ALS) classification, where existing methods have not achieved satisfactory accuracy, I propose an effective and robust transformer-based framework called $SF^2Former$. Leveraging the vision transformer (ViT) concept, $SF^2Former$ employs a novel linear fusion of spatial and frequency domain information to efficiently extract robust local and global discriminative features. This study pioneers in applying a transformer-based deep model for ALS classification, achieving state-of-the-art performance compared to existing popular ML methods.

Finally, a new perspective in solving the domain shift issue for MRI data is designed by identifying and addressing the dominant factor causing heterogeneity within the dataset. An unsupervised DA method called DAMS (Domain Adaptation of MRI Scanners) is developed to align domain-invariant features between source and target domains by minimizing discrepancies in their feature maps. Instead of treating the entire dataset as a single source or target domain, the method processes data based on the primary factor driving variations. Furthermore, my research extends the concept of handling domain shift through black-box source-free domain adaptation (SFDA), which aggregates knowledge from multiple source domains and eliminates the need to access source data during target domain adaptation. This thesis offers innovative solutions to domain shift challenges in MRI data analysis, benefiting researchers not only in medical imaging fields but also in computer vision.

Preface

All methods presented in this thesis are published at prominent venues of computer vision and medical imaging. I am accountable for conceptualizing and executing algorithm designs, experiments, and the majority of paper writing for all projects.

Chapter 3: Rafsanjany Kushol, Alan H. Wilman, Sanjay Kalra, and Yee-Hong Yang. "Dsmri: Domain shift analyzer for multi-center mri datasets." Diagnostics 13, no. 18 (2023): 2947.

Chapter 4: Rafsanjany Kushol, Sanjay Kalra, and Yee-Hong Yang. "DeepDSMRI: Deep Domain Shift analyzer for MRI." In Conference on Medical Image Understanding and Analysis, Accepted. Cham: Springer Nature Switzerland, 2024.

Chapter 5: Rafsanjany Kushol, Pedram Parnianpour, Alan H. Wilman, Sanjay Kalra, and Yee-Hong Yang. "Effects of MRI scanner manufacturers in classification tasks with deep learning models." Scientific Reports 13, no. 1 (2023): 16791.

Chapter 5: Rafsanjany Kushol, Abbas Masoumzadeh, Dong Huo, Sanjay Kalra, and Yee-Hong Yang. "Addformer: Alzheimer's disease detection from structural mri using fusion transformer." In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE, 2022.

Chapter 6: Rafsanjany Kushol, Collin C. Luk, Avyarthana Dey, Michael Benatar, Hannah Briemberg, Annie Dionne et al. "SF2Former: Amyotrophic lateral sclerosis identification from multi-center MRI data using spatial and frequency fusion transformer." Computerized medical imaging and graphics 108 (2023): 102279.

Chapter 7: Rafsanjany Kushol, Richard Frayne, Simon J. Graham, Alan H. Wilman, Sanjay Kalra, and Yee-Hong Yang. "Domain adaptation of mri scanners as

an alternative to mri harmonization." In MICCAI Workshop on Domain Adaptation and Representation Transfer, pp. 1-11. Cham: Springer Nature Switzerland, 2023.

This thesis is a concatenation of the above six papers.

Acknowledgements

In the name of Allah, the Most Gracious, the Most Merciful. All praise and thanks are due to Allah, the Lord of all the worlds, for granting me the strength, patience, and guidance which has illuminated every step of my academic journey. Without His unwavering support, none of this would have been possible.

I extend my sincere appreciation and gratitude to my esteemed supervisors Prof. Herbert Yang and Dr. Sanjay Kalra, whose expertise, encouragement, and constructive feedback have been invaluable throughout this research endeavour. Their mentorship has not only shaped the trajectory of this thesis but also profoundly influenced my scholarly growth.

I am thankful to the members of my thesis committee, Dr. Alan Wilman and Prof. Pierre Boulanger, for their insightful suggestions and scholarly guidance. Their collective wisdom and expertise have significantly enriched the quality of this work.

I extend my sincere appreciation to the Faculty of Graduate Studies and Research, as well as the Department of Computing Science, for their support over the past five years. I am thankful to all the staff members who have assisted me along the way. Particularly, I am grateful to my labmates at the Computer Graphics and ALSNIRU Labs, including Dong Huo, Abbas Masoumzadeh, G.M. Mashrur E Elahi, H M Ata-E-Rabbi, Pedram Parnianpour, Collin C. Luk, and Avyarthana Dey for their invaluable contributions to this thesis.

I would like to express my deepest gratitude to my beloved family, whose unwavering love, support, and understanding have been my greatest source of strength throughout this demanding journey. To my parents, I can never repay the endless

prayers and support you have given me. A heartfelt thanks to my amazing wife whose tireless care and dedication in managing our home allowed me to concentrate fully on my studies. I am also incredibly fortunate to have five blessed children, whose patience and cooperation made this journey not only possible but meaningful. Your sacrifices and steadfast belief in me have truly been the foundation of my success.

I am immensely grateful to my cherished friends for their constant encouragement, camaraderie, and moral support. Their presence has been a source of strength and inspiration, making the challenges of academia more manageable.

Finally, I acknowledge with gratitude the financial support provided by the Prime Minister Fellowship Bangladesh, Canadian Institutes of Health Research, ALS Society of Canada, Brain Canada Foundation, and Natural Sciences and Engineering Research Council of Canada, which enabled me to pursue my doctoral studies.

Last but not least, I extend my deepest gratitude to all those whose names may not be mentioned here but whose contributions, whether big or small, have left an indelible mark on this thesis and my academic growth. May Allah bless you all abundantly.

Table of Contents

List of Tables

List of Figures

Abbreviations

ABIDE Autism brain imaging data exchange.

AD Alzheimer's disease.

ADNI Alzheimer's disease neuroimaging initiative.

AIBL Australian imaging, biomarker and lifestyle.

ALS Amyotrophic lateral sclerosis.

ALSFRS-R Revised amyotrophic lateral sclerosis functional rating scale.

CALSNIC Canadian ALS neuroimaging consortium.

CNNs Convolutional neural networks.

CSF Cerebrospinal fluid.

CST Corticospinal tract.

DL Deep learning.

DTI Diffusion tensor imaging.

DWI Diffusion weighted imaging.

FLAIR Fluid attenuated inversion recovery.

fMRI Functional MRI.

GLCM Gray level cooccurrence matrix.

LMN Lower motor neurons.

MCI Mild cognitive impairment.

ML Machine learning.

MLP Multilayer perceptron.

MR Magnetic resonance.

MRI Magnetic resonance imaging.

PCG Precentral gyrus.

PPMI Parkinson's progression markers initiative.

RF Random forest.

SFDA Source-free domain adaptation.

SVM Support vector machine.

t-SNE t-distributed stochastic neighbor embedding.

UDA Unsupervised domain adaptation.

UMAP Uniform manifold approximation and projection.

UMN Upper motor neurons.

ViT Vision transformer.

Chapter 1

Introduction

1.1 Background and Motivation

Machine learning (ML) is a mathematical method based on statistics by which a computer model is created to perform specific tasks by learning from existing data and has been applied in clinical applications for many years [1–3]. A prominent ML branch known as deep learning (DL) builds models using layers of interconnected neurons to learn critical insights from existing data and to predict the outcome of new data. Unlike traditional ML methods, DL networks automate feature extraction and selection, making them user-friendly and more prevalent than classical ML techniques. Recent research has demonstrated that DL, particularly convolutional neural networks (CNNs), are an effective strategy for classifying, segmenting, and detecting objects of interest in medical images [4–6]. CNNs, as statistical tools, learn the input data's statistics under the assumption of identical independent distribution (IID). Under this assumption, a trained CNN model is expected to perform consistently on samples with similar or identical distributions. Hence, the practical efficacy of DL frameworks depends on their successful generalization to unknown datasets [7].

Magnetic resonance imaging (MRI) is a versatile, non-invasive imaging modality offering exceptional contrast for analyzing soft tissue. MR images have useful applications, including diagnostics, due to the varied appearance of organs, tissues, and pathology. High-resolution images of brain anatomy are obtained during MRI scans,

allowing medical professionals and researchers to examine different neuroanatomical aspects like cerebral structures, white matter, and grey matter. Radiologists and neurologists use MRI scans to identify abnormalities, such as tumors, lesions, atrophy, or other anatomical changes, that may be signs of disorders like multiple sclerosis, epilepsy, and brain tumors [8–10]. Training a DL model requires sufficient data (e.g., MR images, clinical scores) and/or their corresponding ground truth. The network uses the training data to adjust its internal parameters (up to many millions). The robustness of the model is highly dependent on the inclusion of a large number of relevant samples in the training phase. During deployment, the trained model is applied to unseen samples, leveraging its learned parameters to formulate predictions. While numerous earlier studies have evaluated various DL models across diverse MRI datasets, the generalization issue of DL models on MR images remains [11].

In medical research and clinical applications, the utilization of MRI data sourced from multiple centers has become increasingly prevalent [12–14]. Large-scale multi-center MRI datasets play a vital role in advancing medical research, not only in aiding in the understanding, diagnosing, and treating of a wide range of disorders but also in training DL models. However, the inherent variability among these centers presents challenges due to a phenomenon known as "domain shift", which can impact the quality and reliability of the analysis. Domain shift is the term used to describe the variances in data distributions across different centers resulting from variations in hardware, acquisition protocols, patient demographics, and environmental factors [15]. Several parameters contribute to this phenomenon, including the imaging protocol (comprising aspects such as flip angle, acquisition orientation, slice thickness, and resolution) and the scanner itself (encompassing manufacturer, model, magnetic field strength, and the number of channels per coil). As a result, the appearance, contrast, intensity distribution, spatial resolution, and noise level of MR images differ qualitatively and quantitatively from site to site and study to study [16].

The problem of domain shift creates several challenges in analyzing and interpret-

ing MRI data using ML/DL models. Firstly, it significantly impacts the performance and reliability of ML analysis pipelines, in particular, models trained on data from one center often struggle to generalize effectively when they are applied to data from other centers [17]. This limitation has hindered the widespread adoption of automated tools for tasks like diagnosis, treatment planning, and disease monitoring, as their effectiveness relies on their ability to handle data from diverse sources. Secondly, domain shift can introduce biases and confounds in research studies that utilize multi-site MRI datasets [18]. In the context of clinical trials or population studies involving data from multiple centers, the variations originating from domain shift might distort statistical analysis, leading to erroneous conclusions and misleading findings. Thirdly, the inherent variability in scanner hardware and software across centers can introduce technical discrepancies, further complicating the comparison and fusion of data. These issues pose significant challenges for researchers and clinicians seeking to extract reliable and reproducible insights from multi-center MRI datasets. Effectively addressing the challenges associated with domain shift in multi-site MRI datasets requires advanced techniques and methodologies.

Domain adaptation (DA) and harmonization methods [19–21] aim to bridge the gap among different domains by aligning and harmonizing the data from different centers. The primary goal of DA is to make ML models perform well in the target domain, even when they are trained with data from a different source domain. This is crucial because, in many real-world scenarios, it is impractical or expensive to collect labeled data for the target domain. As depicted in Fig. 1.1, the domain shift phenomenon, marked by discrepancies in data distributions, highlights the significance of DA. The top row of Fig. 1.1 illustrates how variations in image acquisition protocol can lead to significant discrepancies in imaging characteristics, a common challenge in fields like retinal and MR images. In contrast, the bottom row exemplifies the purpose of DA: mitigating the distribution disparities between the source and target domains through adaptation techniques. However, prior to developing DA or harmonization

algorithms, it is crucial to gain a comprehensive understanding of the nature and extent of domain shift present in both source and target datasets. Unfortunately, the lack of adequate tools for domain shift analysis serves as a significant bottleneck, hindering the development and validation of DA techniques.

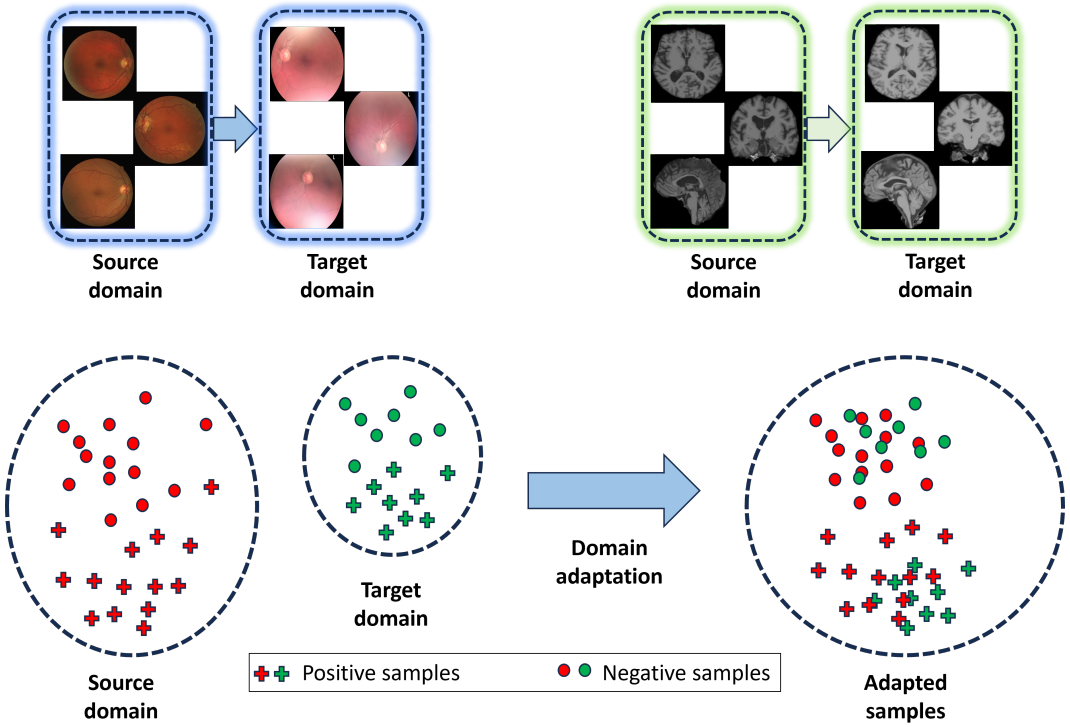


Figure 1.1: Illustration of the "domain shift" phenomenon with significant contrast variation between source and target domain (top row) and the fundamental of domain adaptation (distribution of source and target samples before and after adaptation).

Although many research studies have demonstrated exemplary performance on specific domains and MRI protocols, the applicability of these methods in the target data with different imaging distributions remains questionable [22]. To ensure that the trained models can be used effectively in real-world clinical practice, it is essential to overcome the aforementioned challenges posed by domain shift. Towards this, some studies [23–25] used supervised domain adaptation (SDA) techniques that require labels from both source and target domains. Some researchers [26, 27] trained their models on source domains and fine-tuned the trained models with partially labeled data from the target domain in a semi-supervised fashion. Unsupervised domain

adaptation (UDA) methods [28–32] do not require the ground truth information from the target domain but require accessing the labeled source data during the training process. Taking a step further towards more realistic configuration, source-free domain adaptation (SFDA) approaches [33, 34] eliminate the need for direct access to source data during model training. As shown in Fig. 1.2, the most notable difference between UDA and SFDA is that the UDA models leverage raw source and target domain data, whereas SFDA methods only utilize the trained parameters or predictions of the source domain and then adapt them with unlabeled target data. However, existing techniques do not analyze the nature of heterogeneity present within the source/target domain and process the entire dataset to mitigate the effects of domain shift.

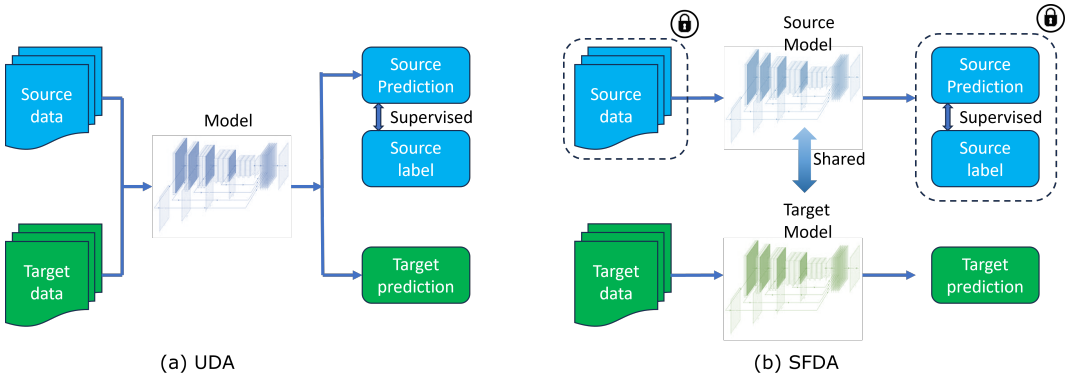


Figure 1.2: The basic comparison of (a) UDA and (b) SFDA configurations.

1.2 Contributions

Motivated by the aforementioned challenges and limitations of the existing literature, firstly, I propose a novel *Domain Shift analyzer for MRI* (DSMRI) framework to comprehensively assess the extent of domain shift within MRI datasets in Chapter 3. The proposed framework provides key insights into domain shift factors and identifies the dominant contributors to data heterogeneity. DSMRI integrates knowledge from diverse domains, including spatial, frequency, wavelet, and texture analysis. This multi-domain feature extraction approach strengthens the framework's ability

to capture various aspects of domain shift. Moreover, deriving the unique features from the frequency domain to capture low and high-frequency image information and incorporating wavelet domain features to measure sparsity and energy within wavelet coefficients significantly enhance the robustness of domain shift analysis. Furthermore, using visualization techniques such as t-SNE [35] and UMAP [36] enrich the framework's ability to visually represent and interpret domain shift effects. The effectiveness of the proposed approach is demonstrated using structural MRI data (*e.g.*, T1-weighted, T2-weighted, and FLAIR) across seven large-scale multi-site neuroimaging datasets.

In Chapter 4, I introduce another unsupervised framework called *Deep Domain Shift analyzer for MRI* (DeepDSMRI), which analyzes domain shift in MRI data using various deep models pre-trained on the ImageNet dataset [37] without requiring additional training on MRI data. The proposed framework demonstrates its efficacy in determining domain shift not only in structural MRI (*e.g.*, T1-weighted, T2-weighted, and FLAIR) but also in other advanced MRI modalities such as diffusion-weighted imaging (DWI), and functional MRI (fMRI). To the best of my knowledge, this is the first work to analyze and quantify domain shift in multi-modal MRI data using deep learning.

In Chapter 5, the focus is on the non-biological factors of variability in neuroimaging data, which pose a barrier to the practical applications of DL algorithms in the medical domain. Specifically, I investigate the impact of scanner vendor variability on various disease classification tasks across multiple DL models. My analysis reveals a significant decline in classification accuracy when DL models are tested with data from different scanner manufacturers. Subsequently, experiments show that employing a popular statistical harmonization technique called ComBat fails to provide improvements in disease classification performance when it is applied to multi-center datasets of 3D structural MR images.

In my previous investigation of disease classification performance using existing

deep models, I discovered that no existing models provide satisfactory accuracy for the challenging task of ALS classification. To address this, I propose an effective and robust transformer-based framework called $SF^2Former$ (spatial and frequency fusion transformer) for ALS classification using multi-modal brain imaging data in Chapter 6. Leveraging the concept of the vision transformer (ViT) [38], $SF^2Former$ distinguishes ALS samples from healthy controls by utilizing a set of intermediate 2D coronal slices extracted from 3D MRI data. The framework employs a novel linear fusion of spatial and frequency domain information to efficiently extract robust local and global discriminative features. To the best of my knowledge, this is the first study to apply a transformer-based deep model for ALS classification, achieving state-of-the-art performance compared to many popular CNN-based DL methods. Notably, $SF^2Former$ not only excels in ALS classification but also demonstrates superior performance in classifying other neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease.

In Chapter 7, I propose a new perspective in solving the domain shift issue for MRI data by identifying and addressing the dominant factor causing heterogeneity in the dataset. Specifically, I design a multi-source UDA method which aligns the domain-invariant features of the source and target domain by minimizing the discrepancy between these domains. I combine maximum mean discrepancy (MMD) [39] and a modified correlation alignment (CORAL) [40] loss functions to extract pairwise domain-specific invariant features. Instead of regarding the entire dataset as a source or target domain, the dataset is processed based on the dominant factor driving data variations which is the scanner manufacturer.

Finally, I extend the concept of handling domain shift under the formulation of black-box SFDA, which eliminates the need for concurrent access to both source and target data during training. My proposed technique employs self-distillation and consistency regularization for pseudo label generation and refinement with a learnable ensemble network. The successful implementation of these proposed strategies has

significant implications for the fields of computer vision and medical imaging. It enables the transfer of knowledge from one or multiple domains while preserving privacy and facilitates the application of learned knowledge to domains where data scarcity is a prevailing challenge.

To assess the effectiveness of the proposed frameworks, I conduct comprehensive experiments with a wide range of diverse multi-center MRI datasets, including participants with amyotrophic lateral sclerosis (ALS), Alzheimer's disease (AD), Parkinson's disease (PD) and autism spectrum disorder (ASD) in addition to healthy controls (HC). To foster reproducibility and knowledge sharing, the Python source codes of the proposed frameworks have been made publicly available.

1.3 Organization

The rest of this dissertation is organized as follows. Chapter 2 reviews existing literature on domain shift in MRI data, domain shift analysis tools, methods for ALS and AD classification, MRI harmonization, and domain adaptation techniques. Two prominent proposed methods to analyze domain shift on MRI data are introduced in Chapters 3 and 4. The effects of different scanner manufacturers on various disease classification tasks are discussed in Chapter 5. Afterwards, a novel disease classification framework is presented in Chapter 6. Finally, Chapter 7 covers DA techniques designed to handle domain shift in MRI data. For each proposed framework, I first describe the approach and then present the experimental results. Chapter 8 concludes the dissertation and discusses future research directions. Appendices A and B provide additional experimental details and results related to the corresponding chapters.

Chapter 2

Related Works

2.1 Domain Shift in MRI Data

Prior studies have widely acknowledged and examined the presence of domain shift in diverse MRI datasets. Researchers have consistently reported variations and challenges originating from domain shift, highlighting the need for robust analysis techniques [41–43].

A study by Dadar *et al.* [44] examined the impact of scanner manufacturers on a brain MRI dataset collected from multiple imaging centers. They reported significant differences in gray and white matter volume estimation among scanner manufacturers. These variations affected the reliability of automated brain segmentation algorithms, resulting in inconsistent outcomes from different centers. In another investigation by Tian *et al.* [45], domain shift effects were analyzed to reduce the site effects on gray matter volume maps using a travelling-subject MRI dataset obtained from various sites. They considered several underlying domain shift factors such as scanner manufacturer, model, phase encoding direction, and channels per coil. Interestingly, they found that the scanner manufacturer is the most significant parameter causing domain shift, followed by the scanner model.

In another study, Lee *et al.* [46] explored the effects of changing MRI scanners on whole-brain volume change estimation at different time point visits. They identified that inter-vendor (e.g., Philips to Siemens) scanner changes led to more significant

effects on percentage brain volume change than intra-vendor (e.g., GE Signa Excite to GE Signa HDx) scanner upgrades. Additionally, Glocker *et al.* [47] conducted an empirical study to investigate the impact of scanner effects when using ML on multi-site neuroimaging data. The authors discovered that even after meticulous pre-processing using advanced neuroimaging tools, a classifier could identify the origin of the data (e.g., scanner) with high accuracy. Furthermore, Panman *et al.* [48] experimented with 8-channel and 32-channel head coil configurations using structural, diffusion and functional MR images while keeping all other image acquisition parameters identical. They showed that the variations in the number of head coils could considerably impact the outcomes of analysis methods, despite having the other acquisition parameters synchronized.

The above studies collectively highlight the pervasive presence of domain shift in multi-center MRI datasets. The observed variations in image characteristics and acquisition parameters across centers pose significant challenges for analysis and interpretation.

2.2 Quality Assessment Methods for MRI Data

MRIQC [49] is an open-source tool developed to automatically predict the quality of MRI data acquired from unseen sites, as manual inspection is subjective and impractical for large-scale datasets. The tool extracts a set of spatial domain features to train an ML classifier and predict whether a scan should be accepted or excluded from the analysis. The authors validated that MRIQC accurately predicted image quality on an unseen dataset of multiple scanners and sites with approximately 76% accuracy. To address the errors and inconsistencies in brain image segmentation, Mindcontrol [50], a web-based application, was designed to allow a user to inspect brain segmentation data and manually correct errors visually. The user can view and interact with 3D brain images, including the ability to adjust opacity, slice orientation, and zoom level for data curation and quality control (QC).

Osadebey *et al.* [51] presented a quality metric scheme for structural MRI data in multi-site neuroimaging studies. Their system evaluates image quality based on factors such as luminance contrast, texture analysis, and lightness and generates a total quality score. The authors demonstrated their framework's effectiveness by applying it to large-scale multi-center MRI data and concluded that it correlates well with human visual judgment. The quality evaluation using multi-directional filters for MRI (QEMDIM) [52] is a technique which is capable of detecting various distortions, including Gaussian noise and motion artifacts. Their method utilized mean-subtracted contrast-normalized (MSCN) coefficients to extract image statistics in the spatial domain and achieved satisfactory accuracy in identifying low-quality images affected by different artifacts or noises compared to undistorted images.

In another study, Esteban *et al.* [53] proposed a crowdsourcing approach for collecting MRI quality metrics and expert quality annotations to train both humans and machines in assessing the quality of MRI data. They revealed that the ML algorithms trained on the crowdsourced data perform comparably to human raters in evaluating image quality. The strategy developed by Oszust *et al.* [54], NOMRIQA, used high-boost filtering to intensify the high-frequency points, which allows the identification of various distortions. Their method utilized the fast retina key-point descriptor and the support vector regression classifier to generate a quality score, which assists in detecting distorted T2-weighted images.

Bottani *et al.* [55] introduced an automated QC method for brain T1-weighted MRI in a clinical data warehouse. Their technique involves extracting spatial domain features using a convolutional neural network (CNN) to predict scans which need to be excluded. They showed that their method could recognize images with potential quality issues, such as artifacts or motion-related distortions, and detect acquisitions for which gadolinium was injected. Lastly, an overview of various no-reference image quality assessment (NR-IQA) methods designed explicitly for MRI data can be found here [56]. The authors discussed the challenges associated with evaluating MRI image

quality due to the complex and dynamic nature of MRI data, including the influence of various acquisition parameters, image artifacts, and population-related factors.

These QC studies focus mainly on automatically detecting artifacts or poor-quality samples to reduce manual effort and decide whether a particular scan should be accepted or excluded from the analysis. These studies neither emphasize quantifying the degree of domain shift from these QC features nor analyze which features are correlated to domain shift.

2.3 Existing Domain Shift Analysis Tools

The tools introduced by Sadri *et al.*, MRQy [57] and Guan *et al.*, DomainATM [20] can be considered two closest studies to analyze domain shift on MRI data. MRQy is mainly designed for the QC of MRI data by which manual effort to filter poor-quality data can be automated for clinical research studies. It uses different spatial domain image quality-related metrics to address different types of noise, shading, inhomogeneity, and motion artifacts. Although they provided an example of detecting site effects using their proposed features, the experimentation on large-scale datasets with more scanner/acquisition protocol variations revealed that MRQy features could not cluster the data accurately. Secondly, MRQy used metadata such as image/voxel dimension from the file header. These features become identical for all the center's data after commonly used preprocessing steps like skull stripping or registration; hence, they are not fruitful for site effect analysis.

On the other hand, DomainATM offers visualization of data distribution as well as measures the domain shift distance for the original or synthetic data. Then, they implemented some classical DA methods to show the effectiveness of these methods in reducing the domain shift. However, their tool cannot take raw neuroimaging data, such as NIfTI files, directly as input. To analyze real-world data with DomainATM, the user must process the data with Anatomical Automatic Labeling (AAL) atlas and then extract the grey matter volumes for each region of interest (ROI), making the

tool inconvenient for many applications. Most importantly, these grey matter features are not meaningful regarding the domain shift measurement, which is reflected in the experimental section 3.2.8. My proposed frameworks DSMRI and DeepDSMRI are compared with MRQy and DomainATM to demonstrate the strength of the proposed features in analyzing the domain shift in multi-center large-scale MRI datasets.

2.4 ALS Classification

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disorder that typically manifests in individuals in their late fifties to early sixties, affecting both upper and lower motor neurons in the nervous system. The degeneration of upper motor neurons (UMN) leads to symptoms such as spasticity, exaggerated reflexes, and slowness of movement while the degeneration of lower motor neurons (LMN) causes weakness, muscle atrophy, and fasciculations. As the disease progresses, patients may experience loss of limb function, difficulty walking, speaking, and eventually breathing. Respiratory failure often becomes the cause of death. On average the survival time is 3-5 years from symptom onset [58]. The precise pathophysiology underlying neurodegeneration in ALS remains insufficiently understood. While a small percentage (5-10%) of ALS cases are familial, the vast majority of patients (90-95%) present with sporadic forms of the disease [59]. Currently, only a few pharmacologic therapies, such as riluzole and edaravone, have been approved for use in the early stages of the disease to slow progression and improve survival [60]. Unfortunately, no therapies are available that can halt disease progression entirely.

In ALS, neuroimaging is routinely utilized to rule out other diseases but does not play a direct role in making a definitive diagnosis. While structural changes in affected regions such as the precentral gyrus (PCG) and corticospinal tract (CST) can be observed in a minority of cases through visual inspection [61, 62], the majority of MRI scans of ALS patients do not exhibit such changes (Fig. 2.1). Therefore, one significant challenge to developing effective treatments is the absence of estab-

lished biomarkers that can accurately track disease progression or aid in early diagnosis. Recent research efforts have focused on establishing MRI measures as potential biomarkers for tracking disease progression in ALS. The development and demonstration of an automated method that can accurately classify patients from healthy controls is a critical first step towards the eventual inclusion of the technique in the clinical diagnosis of ALS.

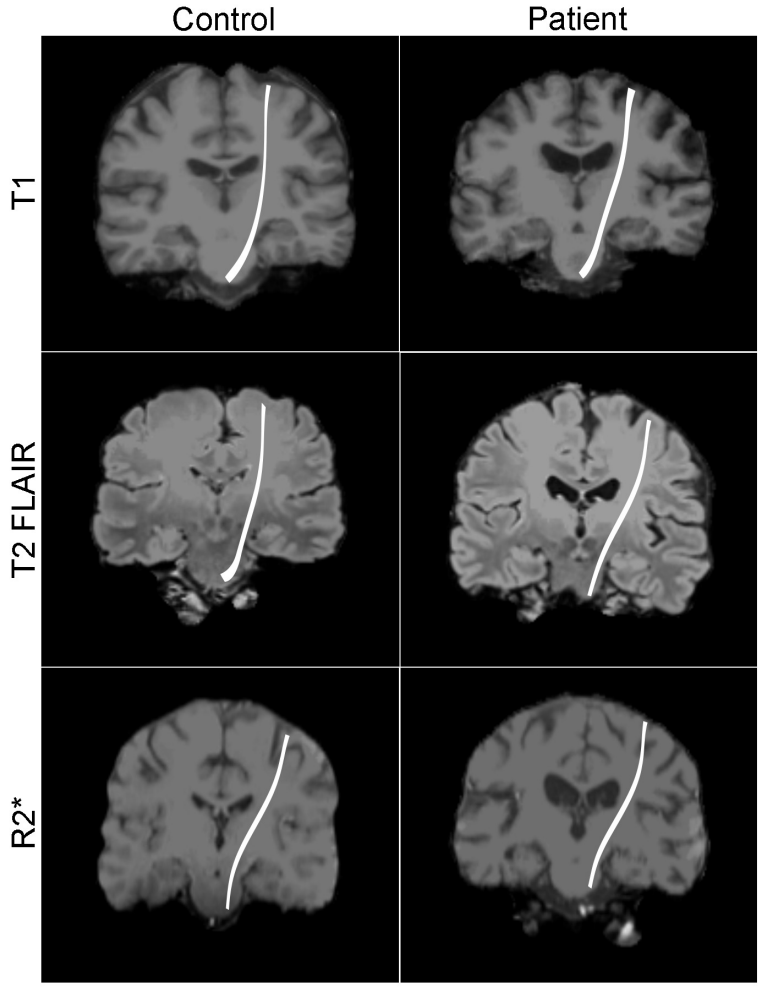


Figure 2.1: MR images of controls and ALS patients for T1-weighted, T2-FLAIR and R2* modalities. Coronal images are sampled at the plane of the precentral gyrus with a white line demonstrating the approximate path of the corticospinal tract within each plane. There are no visually discernible features in the gray and white matter between controls and ALS patients.

The involvement of iron accumulation in the motor cortex area has been reported in multiple in vivo and ex vivo studies conducted on ALS cohorts. When comparing

different MRI sequences such as T1-weighted, T2*-weighted, and FLAIR, the T2*-weighted sequence has been shown to better capture hypointensities in the precentral gyrus gray matter (PGGM) in ALS patients [63]. In one study, Wang *et al.* [64] discovered increased R2* values in the primary motor cortex in ALS compared to healthy controls. Furthermore, Hecht *et al.* [65] revealed the presence of more hyperintense signals in the CST in FLAIR images compared to T1-weighted, T2-weighted, and proton density-weighted images. Similar findings were also demonstrated by Jin *et al.* [66] with increased CST hyperintensity in ALS compared to control samples in the juxtacortical PCG. Another study by Fabes *et al.* [67] highlighted the significance of FLAIR intensity in the CST and corpus callosum of ALS patients in comparison to normal controls. Additionally, Li *et al.* [68] demonstrated a considerable decrease in fractional anisotropy (FA) in the left CST for ALS cohorts using structural and diffusion tensor MRI. Lastly, Alberich *et al.* [69] provided an extensive overview of imaging biomarkers in ALS. These crucial neuroimaging findings motivate my investigation into different MRI measures for the automatic differentiation between ALS patients and healthy controls.

VoxelHop was proposed by Liu *et al.* [70], which used T2-weighted structural MR images to detect ALS. However, their evaluation was limited to a small-scale dataset consisting of 20 controls and 26 patients. By utilizing recurrent neural networks and random forest classifiers, Thome *et al.* [71] designed a feature set from structural and functional resting-state MRI which was able to achieve a maximum classification accuracy of 66%. On the other hand, Elahi *et al.* [13] introduced a modified co-occurrence histogram of oriented gradients (M-CoHOG) method for feature selection using 2D coronal slices of T1-weighted images. While their technique achieved 76% classification accuracy in a single-center dataset, it exhibited poor consistency when applied to an extended version of the multi-center database. On top of that, M-CoHOG required laborious efforts from experts to manually select the appropriate coronal slices for each individual. Moreover, Chen *et al.* [72] employed FA informa-

tion from diffusion tensor imaging (DTI) and a linear kernel support vector machine (SVM) to classify ALS from healthy controls and obtained a classification accuracy of 83%. However, their dataset was also limited, comprising 22 ALS patients and 26 healthy subjects. In another study by Kocar *et al.* [73], using DTI and texture analysis with a linear SVM classifier, the authors reported approximately 80% classification sensitivity and specificity.

2.5 Alzheimer's Disease Classification

Alzheimer's disease (AD) is a gradual fatal condition of the brain affecting 1 in 10 people above the age of 65 that deliberately impairs memory and thinking skills. The aetiology of the disorder is not adequately understood. However, genetics and environmental factors are thought to be involved [74]. The disease grows steadily as irregular protein fragments named plaques and tangles are accumulated in the brain and destroy brain cells. They originate from the hippocampus region of the brain where memories are first developed. Afterwards, more plaques and tangles expand into different areas of the brain and compromised brain functionalities by continuing killing neurons [75]. This spreading around the brain basically causes the distinct stages of AD such as mild cognitive impairment (MCI). Some drugs are generally used for the treatment of AD which promotes healthy cognition and memory; however, they do not stop neurodegeneration. As a result, there is no medication to cure the disease.

In the case of AD compared to healthy control mostly the region known as the hippocampus is affected at the beginning stage and considered as the prominent feature to classify AD [76]. In later stages, the ventricle and cortex area also got affected. Figure 2.2 shows a coronal slice of AD patient and healthy control highlighting left and right hippocampus (yellow box) and ventricle (green box) regions. Existing research, which has been conducted to predict and classify AD using diverse ML/DL strategies, can be roughly categorized into the following four types:

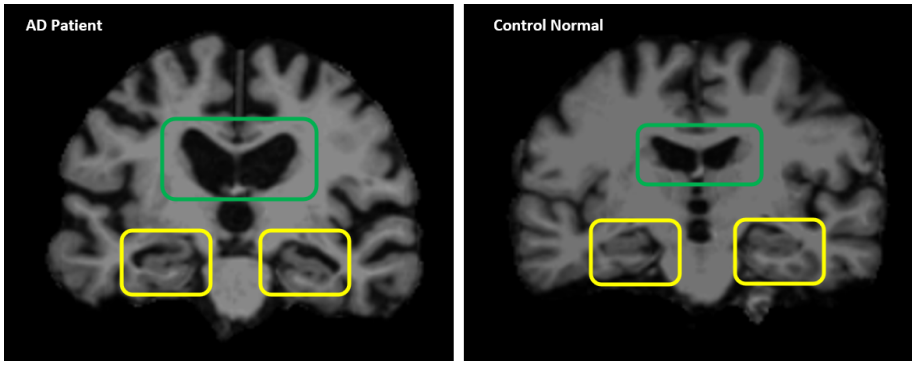


Figure 2.2: Distinguishable regions of the brain related to AD in structural MRI. It illustrates a coronal slice of AD patient and healthy control highlighting left and right hippocampus (yellow box) and ventricle (green box) regions.

Voxel-based: This is the most automated and straightforward technique by which the full 3D MRI scan is used for feature extraction. However, these methods require high computational cost, memory, and feature dimensionality. On top of that, a large portion of the volume does not contribute any distinguishing feature in identifying AD. A 3D CNN is designed by Shahamat *et al.* [77] utilizing a genetic algorithm-based brain masking (GABM) to capture features related to AD. In another work, Solano *et al.* [78] proposed a cost-efficient 3D DenseNet-121 architecture to classify different stages of AD.

Slice-based: Several research works develop their methods based on axial projection view, whereas few studies are also based on the coronal view. The slice-based schemes reduce the computational complexity as well as make it easy to leverage transfer learning models trained on computer vision-related 2D datasets. Ebrahimi *et al.* [79] introduced a deep sequence-based network to classify AD given that the inputs are coming from manually chosen 2D coronal slices. In another work, Zhang *et al.* [80] combined gray matter slice region and attention mechanism to classify AD and MCI using T1-weighted structural MRI.

ROI-based: This approach involves adequate manual observation, prior knowledge, and/or experts support to ensure the right region is processed through the network. In one study, Lin *et al.* [81] derived AD related features with the help of FreeSurfer

[82] which includes cortical thickness, surface area, and cortical volume. Moreover, Liu *et al.* [83] performed joint hippocampus segmentation and AD classification using a 3D DenseNet-based architecture.

Patch-based: The key obstacle in the patch-based procedure is determining meaningful image patches to extract local and global features. Qiu *et al.* [4] integrated multi-modal inputs like MRI and neuropsychological testing scores in a CNN model to classify AD. In another study, Li *et al.* [84] utilized the DenseNet architecture with different 3D image patches clustered using K-Means clustering to distinguish AD from healthy controls.

2.6 Transformer in Medical Image Analysis

The transformer architecture [85] was initially introduced in the context of natural image processing (NLP). It allows for the capturing of long-term dependencies as well as parallel processing of multiple words or patches. However, its application in computer vision has been hindered by its high computational requirements. To overcome this limitation, the vision transformer (ViT) [38] employed non-overlapping patch tokens to embed an image, thereby reducing the spatial dimension of the representation. ViT has achieved state-of-the-art performance on the renowned ImageNet dataset [37] for image classification. One of the drawbacks of ViT is its incapability to learn dependencies within the patch. The Swin transformer [86] addressed this limitation by leveraging a hierarchical structure that captures both local to global relationships. Moreover, the global filter network (GFNet) [87] has been proposed for capturing both long-term and short-term spatial relationships in the Fourier domain. By applying a discrete Fourier transform with a global convolution, the GFNet reconstructs ViT's self-attention layer, resulting in considerable performance improvements. Another study by Touvron *et al.* [88] introduced data-efficient image transformers (DeiT) that utilized knowledge distillation, allowing ViT to perform well even on smaller datasets.

Transformer-based approaches are not only leading in computer vision tasks but have also demonstrated successful applications in diverse medical image analysis contexts. For instance, TransUNet [89] employed CNNs to extract features, which are then fed into a ViT network for efficient medical image segmentation. TransFuse [90] also leveraged the fusion of ViT and CNN features for various 2D and 3D medical image segmentation tasks. In contrast, MedT [91], which is based on axial-attention, investigated the viability of using transformers without large-scale datasets. Coarse to fine vision transformer (C2FViT) [92] was developed for 3D affine medical image registration using ViT and a multi-resolution strategy. Utilizing the effectiveness of ViT, ScoreNet [93] has been proposed for histopathological image classification, whereas Uni4Eye [94] has been developed for robust ophthalmic image classification. Furthermore, SphereMorph [95], a robust diffeomorphic cortical surface registration network, used a UNet-style architecture and a modified spatial transformer layer. The success of these models demonstrates the enormous promise of transformers in medical image analysis.

2.7 MRI Harmonization Techniques

Harmonization is a technique used to mitigate variations arising from diverse image acquisition protocols. Many efforts have been made to address the negative impact of scanner bias using MRI harmonization, which aims to mitigate site effects while retaining the statistical power to detect biological factors in images. Image translation-based methods like CycleGAN [96] or neural style transfer [97] render harmonized images to address the issue of domain shift. Statistical techniques, such as ComBat [98], have also been utilized to harmonize region of interest (ROI)-extracted biomarkers and alleviate scanner bias.

Existing research shows that ComBat is highly successful in neuroimaging data harmonization, focusing on removing scanner effects from a set of imaging features such as cortical thickness, surface area, and subcortical volumes [98–101]. Pomponio

et al. [102] applied a modified ComBat method to 145 anatomical ROI volumes to eliminate location and scale effects for each ROI. Another study by Horng *et al.* [103] reported better performance by employing radiomic features from lung computed tomography (CT) images with a modified ComBat method.

Nevertheless, the application of a ComBat-based strategy to full-size 3D (NIFTI) images, rather than specific ROIs or extracted features, presents an ongoing challenge. For extensive high-resolution image datasets, memory allocation constraints may impede program execution. Additionally, ComBat-based strategies require some demographic data to be available for all samples, such as sex, age, and disease status, which are aimed to preserve during harmonization. Importantly, adding a new sample to an existing dataset imposes another concern: the need to rerun the entire harmonization process with the newly added data. Furthermore, modifying pixel intensities before training may not be ideal for medical applications, as it may remove meaningful pixel-level details needed for various tasks such as anomaly detection. However, a recent study [21] revealed that existing image translation or statistical approaches including ComBat failed to harmonize cortical thickness from multi-scanner MRI data properly. Unlike existing harmonization methods, which apply transformation to images to reduce scanner bias, my proposed study aims to adapt datasets from various scanners without changing the actual image content.

2.8 Domain Adaptation Methods

In recent years, there is a notable surge in research utilizing diverse DA techniques, including supervised, semi-supervised, or unsupervised approaches, across both segmentation and classification tasks in the field of medical imaging [23, 25–29, 31, 32]. For instance, Ghafoorian *et al.* [26] utilized transfer learning and reported that without fine-tuning the model using target domain data, the pre-trained model failed for brain white matter hyperintensity (WMH) segmentation. A mixup strategy-based unsupervised domain adaptation (UDA) [31] for knee tissue segmentation revealed that

the model trained from scratch with fewer samples lacked generalization and performed worse when tested on different domains. Another UDA approach by Orbes *et al.* [32] used a paired consistency loss to control the adaptation for WMH segmentation. However, none of these studies considered scenarios involving multiple source or heterogeneous target domains.

Wachinger *et al.* [23] designed a supervised classification model for AD detection by regularizing the multinomial regression employing $l1/l2$ norm. Another supervised domain adaptation (SDA) framework by Wolleb *et al.* [25] ignores scanner bias while focusing on pathology-related features for HC vs. multiple sclerosis (MS) patients classification. With pre-training and fine-tuning phases, Zeng *et al.* [27] utilized federated learning for schizophrenia and major depressive disorder classification tasks. Another SDA study by Dinsdale *et al.* [24] generated scanner invariant representation for the age prediction task. Note that these methods require full or partial ground truth from the target domain, which is often time-consuming and costly in a real-life scenario.

In one study, Wang *et al.* [29] developed a pre-trained classifier using source data and fine-tuned this model to new data, which showed improvement in the AD classification task. By learning a common embedding space for source and target samples, the UDA framework Seg-JDOT [28] was developed for MS lesions segmentation. However, these studies overlooked the presence of heterogeneity within the target domain dataset, specifically the variations introduced by different MRI scanners. Addressing this heterogeneity is crucial for accurate analysis. For a comprehensive understanding of the advancements, key aspects, and pitfalls of various DA techniques in medical image analysis, refer to the review studies by Guan *et al.* [19], Choudhary *et al.* [104], and Kumari *et al.* [105].

2.9 Source-free Domain Adaptation

In conventional DA approaches it is assumed that both source and target domain data can be accessed during the adaptation phase. However, this assumption often does not align with the practicalities of many applications, particularly in the domain of medical imaging. Factors such as privacy, confidentiality, and copyright issues can render raw source domain data inaccessible. Furthermore, storing the entire source dataset on resource-limited devices for training purposes can be a formidable challenge.

To address this practical gap, source-free domain adaptation (SFDA) techniques have drawn significant attention in recent years. Based on the level of access to the source model's parameters and predictions, SFDA methods can be classified into white-box and black-box approaches. In white-box SFDA [106–108], full access to the internal details of the pre-trained source model, including its architecture, parameters, and gradients are available. In contrast, black-box SFDA methods [109–111] only have access to the outputs of the pre-trained source model (i.e., predictions) without any insight into the internal architecture or parameters. For a comprehensive overview of recent advancements in SFDA, particularly in the context of computer vision datasets, review studies by Fang *et al.* [112] and Yu *et al.* [113] provide detailed analyses and summaries of the field.

In one medical imaging study, Yang *et al.* [33] proposed Fourier style mining and contrastive learning to produce source-like images through statistical information of the pre-trained source model for polyp and prostate image segmentation. In another medical image segmentation context, Bateson *et al.* [34] estimated class-ratio priors based on anatomical knowledge and maximized mutual information between target images and their label predictions. Hong *et al.* [114] introduced an SFDA method for cross-modality abdominal multi-organ segmentation using entropy minimization and feature map statistics to guide model adaptation. In addition to entropy mini-

mization, another study [115] proposed the Ring loss to constrain the feature vector norm to preserve the target organ's shape constraints for different medical image segmentation tasks.

For medical image segmentation, further advancements in SFDA methodologies have emerged, such as Fourier visual prompting (FVP) [116]. Their method encouraged a pre-trained model to perform effectively in the target domain by adding a visual prompt to the input target data. Another SFDA framework named SFHarmony [117] modelled the imaging features as a Gaussian mixture model and minimized the distance between source and target features across classification, segmentation and regression tasks. Yu *et al.* [118] proposed prototype-anchored feature alignment utilizing bi-directional transport and contrastive learning for cross-modality medical image segmentation.

It is noteworthy to mention that the aforementioned SFDA-based studies using MRI data often overlooked the intrinsic heterogeneity or domain shift within each domain, treating the entire dataset as either a source or target domain. In contrast, my proposed approach aims to leverage an understanding of the significant factors driving domain shift and align the feature space accordingly, harnessing this prior knowledge to improve adaptation outcomes.

Chapter 3

Domain Shift Analyzer for MRI

The degree of domain shift in an MRI dataset is a problem worth investigating and is the principal focus of this chapter. In particular, I propose a novel framework named DSMRI (Domain Shift analyzer for MRI) to qualitatively and quantitatively determine the degree of domain shift present in an MRI dataset. The proposed framework leverages existing MRI quality-related spatial domain features as well as introduces frequency, wavelet and texture domain features to quantify the degree of domain shift. The source code is available at https://github.com/rkushol/DSMRI.

3.1 Proposed Method

3.1.1 Overview

An overview of the proposed DSMRI framework is shown in Fig. 3.1. The 22 features used in the proposed framework are summarized in Table 3.1. The features are extracted from the foreground of 2D slices of 3D MRI in three different directions, i.e., axial, sagittal, and coronal. MRQy [57] is used to detect the foreground of the MR image. However, Signal-to-Noise Ratio (SNR), Contrast-to-Noise Ratio (CNR) and Coefficient of Joint Variation (CJV) features also involve the background intensity information to measure their corresponding quality score.

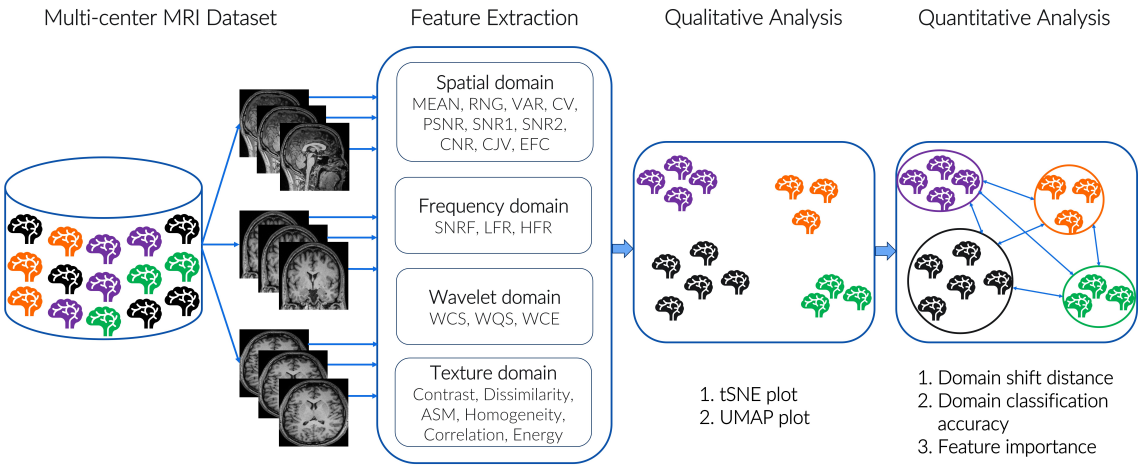


Figure 3.1: An overview of the proposed DSMRI framework. The different colours in the brain icon show that MRI data originated from different sites or may be acquired with distinct image acquisition protocols. Twenty-two significant features are extracted from 2D MRI slices of each subject. Utilizing these feature maps, t-SNE and UMAP methods are used to visualize the position of each scan in a reduced two-dimensional plot. The results are also interpreted in quantitative analysis, where the domain shift distance can be obtained with the maximum mean discrepancy (MMD) distance and the ranking of 22 features to show which features play a more significant role in classifying different domains. Best viewed in color.

3.1.2 Spatial Domain Features

Spatial domain features are based on prior studies [49, 57, 119–121]. Firstly, essential statistical intensity distributions of the foreground (F) are extracted by $MEAN = \frac{1}{HW}\sum_{i,j} F(i,j)$, Range $(RNG) = max(F) - min(F)$, and Variance $(VAR) = \sigma_F^2$, where $H = height, W = width$, and $\sigma^2 = variance$.

Secondly, relevant noise-related features are extracted by incorporating the F and background (B), such as Peak SNR $(PSNR)$ following [57], $SNR1 = \frac{\sigma_F}{\sigma_B}$, $SNR2 = \frac{\mu_{FP}}{\sigma_B}$, and $CNR = \frac{\mu_{FP-BP}}{\sigma_{BP}}$, where, μ = mean, σ = standard deviation (SD), FP = Foreground Patch, and BP = Background Patch. FP and BP are random 5×5 square patches of the F and B, respectively.

Finally, to detect different types of artifacts like shadowing, inhomogeneity, aliasing, and motion, I employ Coefficient of Variation (CV), CJV and Entropy Focus Criterion (EFC) features extracted from the F. The CV is defined as: $CV = \frac{\sigma_F}{\mu_F}$,

Table 3.1: Summary of the proposed features used in our study to quantify the degree of domain shift. [GLCM=Gray-Level Co-occurrence Matrix], [SD=Standard Deviation]

Type	Metric	Description
Spatial domain	MEAN	Mean intensity of the foreground.
	RNG	Intensity range of the foreground.
	VAR	Intensity variance of the foreground.
	CV	Coefficient of Variation to detect shadowing and inhomogeneity artifacts [119].
	PSNR	Peak Signal to Noise Ratio of the foreground.
	SNR1	Signal to Noise Ratio of foreground SD and background SD [57].
	SNR2	Signal to Noise Ratio of foreground patch mean and background SD [49].
	CNR	Contrast to Noise Ratio to detect shadowing and noise artifacts [120]. Higher values indicate better quality.
	CJV	Coefficient of Joint Variation between the foreground and background to detect aliasing and inhomogeneity artifacts [121]. Higher values indicate head motion.
	EFC	Entropy Focus Criterion to detect motion artifacts. Indication of ghosting and blurring induced by head motion [49]. Lower values indicate better quality.
Frequency domain	SNRF	Signal-to-Noise Ratio in the Frequency domain which can be calculated by taking the ratio of the power in the signal to the power in the noise.
	LFR	Low Frequency Response which measures the ability of the MRI scan to capture low-frequency information in the image.
	HFR	High Frequency Response which measures the ability of the MRI scan to capture high-frequency information in the image.
Wavelet domain	WCS	Wavelet Coefficient Sparsity measures the sparse information in the wavelet coefficients, which can indicate the presence of artifacts or inhomogeneities.
	WQS	Wavelet-based Quality Score uses the wavelet transform to analyze the spatial frequency content of the image and calculates a quality score based on the magnitude and phase of the wavelet coefficients.
	WCE	Wavelet Coefficient Energy measures the amount of energy present in wavelet coefficients, which can indicate the presence of artifacts or inhomogeneities.
Texture domain	Contrast	Measures local intensity variations between neighboring pixels. High contrast indicate large intensity differences, while low indicate more uniform regions.
	Dissimilarity	Calculates the average absolute difference between the pixel intensities in the GLCM. It quantifies the amount of local variation in the texture.
	ASM	Angular Second Moment measures the uniformity of the intensity distribution in the image and is often used to describe the texture of the tissue.
	Homogeneity	Measures the closeness of the distribution of elements in the GLCM matrix to the diagonal elements, indicating the level of local homogeneity.
	Correlation	Represents the linear dependency between pixel intensities in the image and measures how correlated the pixels are in a given direction.
	Energy	Reflects the overall uniformity in the image. It is calculated as the sum of the squared elements in the GLCM.

whereas, the CJV can be expressed as: $CJV = \frac{(\sigma_F+\sigma_B)}{|\mu_F-\mu_B|}$. The EFC is defined as: $EFC = \frac{HW}{\sqrt{HW}} log \frac{E}{\sqrt{HW}}$, where E is derived following [57].

3.1.3 Frequency Domain Features

The frequency domain features are calculated after performing a 2D fast Fourier transform (FFT) on F. The FFT transform is performed with the Python `SciPy` library [122].

1. SNRF: The SNR in the Frequency domain assesses signal quality corrupted by noise. It quantifies the power (or energy) ratio in the signal component to the power (or energy) in the noise component in the frequency domain. It can be defined as: $SNRF = 10 * log(\frac{P_{signal}}{P_{noise}})$ where P_{signal} is the power (or energy) in the signal component and P_{noise} is the power (or energy) in the noise component.

2. LFR: The Low Frequency Response (LFR) has the ability to capture low-frequency information in the resulting image. It involves applying a low-pass filter, calculating the amplitude spectrum using the FFT, and measuring the LFR as the square root of the amplitude spectrum. The purpose of the low-pass filter is to attenuate or remove high-frequency components from F, allowing only low-frequency information to pass through. A 3×3 Gaussian filter $[[1,2,1],[2,4,2],[1,2,1]]/16$ is used as a low-pass filter which is convolved with F to obtain a low-pass version of the F. The amplitude spectrum of the low-pass image is then computed using the FFT, representing the distribution of frequencies present in the low-pass image. Finally, the square root operation is performed to linearize the amplitude spectrum and make it more suitable for interpretation. The LFR can be expressed as follows: $LFR = \sqrt{FFT(low\_pass\_image)}$.

3. HFR: Similar to the concept of LFR, the High Frequency Response (HFR) has the ability to capture high-frequency information in an image. Instead of a low-pass filter, a high-pass filter is applied to F, allowing only high-frequency components to pass through while attenuating lower frequencies. This step emphasizes the high-

frequency content of an image. A 3×3 Laplacian filter $[[-1,-1,-1],\ [-1,8,-1],\ [-1,-1,-1]]$ is used as a high-pass filter which is convolved with F to acquire a high-pass version of the F. After applying FFT to the high-pass image, the final HFR can be measured as follows: $HFR = \sqrt{FFT(high\_pass\_image)}$.

3.1.4 Wavelet Domain Features

The wavelet domain features are extracted after performing a 2D discrete wavelet transform (DWT) on F. The wavelet decomposition is implemented with the Python `Pywt` package [123], and some examples of wavelet types include Haar, Daubechies, Discrete Meyer, Symlets, and Coiflets.

1. WCS: The Wavelet Coefficient Sparsity (WCS) is a feature used to measure the amount of sparse information present in the wavelet coefficients of a signal or an image. It quantifies the extent to which the wavelet coefficients are concentrated in a few significant coefficients while the majority are close to zero or negligible. First, a 2D DWT is applied to F, which decomposes an image into different frequency sub-bands, representing different scales or levels of detail. Then, the wavelet coefficients obtained from the wavelet transform are analyzed to determine their sparsity. Various sparsity measurement techniques can be employed, such as counting the number of coefficients above a certain threshold or using sparse representation algorithms like l1-norm minimization. Here, the WCS is measured based on the mean of the coefficients. Coefficients above the mean are considered significant, while those below the mean are considered insignificant. The WCS can be represented as follows: $WCS = \frac{\sum_i^n |Significant\_Coefficient_i|}{n}$.

2. WQS: The Wavelet-based Quality Score (WQS) evaluates the quality of an image by analyzing its spatial frequency content using the wavelet transform. It calculates a quality score based on the magnitude and phase information of the wavelet coefficients. The magnitude represents the strength or energy of each coefficient, while the phase represents the spatial orientation or phase shift. The WQS is calculated

by taking the sum of the product of magnitude and cosine form of the phase for each coefficient. The WQS can be expressed as follows: $WQS = \sum_i^n (magnitude_i * cos(phase_i))$, where, $magnitude_i = |Coefficient_i|$ and $phase_i = angle(Coefficient_i)$.

3. WCE: The Wavelet Coefficient Energy (WCE) measures the amount of energy present in the wavelet coefficients of a signal or an image. It quantifies the overall strength or magnitude of the coefficients, indicating how much information is contained in each coefficient. The energy of each wavelet coefficient is computed by taking the absolute value of its magnitude. The total energy of the wavelet coefficients is then obtained by summing up the energies of all the coefficients. The WCE can be defined as follows: $WCE = \frac{\sum_i^n |Coefficient_i|}{n}$.

3.1.5 Texture Domain Features

Texture features are extracted from the widely used GLCM [124, 125], which represents the spatial relationship between pairs of pixel intensities in an image. These features provide valuable information about the spatial patterns and structures present in an image, enabling the characterization and differentiation of various textures within an image. I employ the Python `scikit-image` package [126], which provides a convenient way to calculate various GLCM texture features. A brief description of six GLCM features employed in the proposed framework is given below:

1. Contrast: This feature measures the local variations or differences in intensity between neighbouring pixels in an image. It provides information about the amount of contrast present in the image texture. It is calculated as the sum of squared intensity differences between neighbouring pixel pairs, weighted by the frequencies in the GLCM matrix. A higher contrast value indicates greater variation or sharp transitions between pixel intensities, representing a more textured or detailed image. Command: `skimage.feature.greycoprops (GLCM, 'contrast')`. Formula:

$$Contrast = \sum_{i,j} (i-j)^2 \cdot GLCM(i,j) \tag{3.1}$$

2. Dissimilarity: This feature calculates the average absolute difference between the pixel intensities in the GLCM. It quantifies the amount of local variation in the texture. Higher values indicate greater pixel dissimilarity. It is similar to contrast but focuses on absolute differences rather than squared differences. Command: `skimage.feature.greycoprops (GLCM, 'dissimilarity')`. Formula:

$$Dissimilarity = \sum_{i,j} |i - j| \cdot GLCM(i, j) \tag{3.2}$$

3. Angular Second Moment (ASM): This feature represents the sum of squared elements in the GLCM matrix and reflects the overall uniformity or homogeneity of the image. A higher ASM value indicates a more homogeneous texture, where the pixel pairs are distributed more evenly across the image. Command: `skimage.feature.greycoprops (GLCM, 'ASM')`. Formula:

$$ASM = \sum_{i,j} (GLCM(i, j))^2 \tag{3.3}$$

4. Homogeneity: This feature measures the closeness of the distribution of elements in the GLCM matrix to the diagonal elements, indicating the level of local homogeneity or similarity in an image's texture. A higher homogeneity value indicates a greater level of similarity between neighbouring pixel pairs in terms of intensity values and spatial relationships. Command: `skimage.feature.greycoprops (GLCM, 'homogeneity')`. Formula:

$$Homogeneity = \sum_{i,j} \frac{GLCM(i, j)}{1 + (i - j)^2} \tag{3.4}$$

5. Correlation: This feature measures the linear dependency between pixel intensities in the image. It indicates how correlated the pixels are in a given direction and provides information about the texture's pattern and organization. A higher value suggests a higher degree of linear correlation between pixel pairs in the image, representing a more organized and patterned texture. Command: `skimage.feature.`

`greycoprops (GLCM, 'correlation')`. Formula:

$$Correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j) \cdot GLCM(i,j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \tag{3.5}$$

6. Energy: This feature reflects the overall uniformity or homogeneity of the image texture. It is calculated by the square root of the sum of squared elements in the GLCM matrix, indicating the concentration or "energy" of pixel pairs with specific intensity values and spatial relationships. A higher energy value suggests a more uniform and homogeneous texture in the image. Command: `skimage.feature.greycoprops (GLCM, 'energy')`. Formula:

$$Energy = \sqrt{\sum_{i,j} (GLCM(i,j))^2} \tag{3.6}$$

3.1.6 Applications

Applications and benefits of analyzing and dealing with domain shift in MRI datasets are numerous. Here are some crucial ones:

1. Improved generalizability: Domain shift analysis facilitates the development of ML models that can generalize across multiple centers. By identifying and mitigating the variations caused by domain shift, the methods become more robust and applicable to data from different imaging centers.

2. Reliable and reproducible research: It helps overcome biases and confounds triggered by the variations across different sites. By accounting for the domain-specific effects, research studies utilizing multi-center MRI datasets can yield more reliable and reproducible results.

3. Cross-center comparison and validation: It enables meaningful comparisons and validation of imaging biomarkers, algorithms, and protocols across various centers. Thus, researchers and clinicians can assess the performance and consistency of imaging techniques and analysis methods in diverse settings.

4. Enhanced collaborative research: Multi-center collaborations have become

prevalent in medical imaging research. Analyzing domain shift encourages data sharing and collaboration among different centers by enabling a harmonized data analysis from various sources. It promotes data integration, pooling, and joint analysis, thereby facilitating large-scale studies and advancing scientific knowledge in the field.

5. Adaptation to new centers and populations: As new imaging centers are established, or new patient cohorts are included in studies, domain shift analysis can guide the adaptation of existing models to these new domain configurations. This reduces the time and effort required to deploy analysis tools in new settings, allowing faster translation of research findings into clinical practice.

6. Quality control (QC) and outlier detection: Analyzing domain shift can serve as a QC measure for MRI datasets. It allows for identifying centers or specific scans that exhibit significant variations compared to others. Such insights can help in data validation as well as detect potential sources of errors or outliers.

3.2 Experiments

3.2.1 Datasets

Seven large-scale multi-center datasets are used in the experimental evaluation of the proposed framework. Publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) [127] and the Australian Imaging, Biomarker and Lifestyle (AIBL) [128] datasets comprise AD patients and HC. The Parkinson's Progression Markers Initiative (PPMI) [129] and the Autism Brain Imaging Data Exchange (ABIDE) [130] are also publicly available datasets containing MRI data with PD and ASD patients. The Canadian ALS Neuroimaging Consortium (CALSNIC) [131] multi-site dataset incorporates ALS patients along with HC. For ADNI and CALSNIC, two independent versions are used, ADNI1/ADNI2 and CALSNIC1/CALSNIC2, respectively. The T1-weighted structural MR images are used for all seven databases. Furthermore, I evaluate the outcomes for the T2-weighted and FLAIR (Fluid Attenuated Inver-

sion Recovery) images of the CALSNIC2 dataset. All the aforementioned datasets comprise data from three widely used scanner manufacturers (GE Healthcare, Philips Medical Systems, and Siemens) except the AIBL, which only includes Siemens vendor data. Tables 3.2 and 3.3 illustrate each dataset's demographics and scanning details, respectively.

Table 3.2: Demographic details of the ADNI1, ADNI2, AIBL, PPMI, ABIDE, CALSNIC1, and CALSNIC2 datasets.

Dataset (#total)	**Group**	**MRI Scanner Manufacturer**					
		GE		**Siemens**		**Philips**	
		Sex (M/F)	**Age (Mean±Std)**	**Sex (M/F)**	**Age (Mean±Std)**	**Sex (M/F)**	**Age (Mean±Std)**
ADNI1	AD	85/85	75.5±7.7	85/85	75.0±7.2	43/33	75.7±7.0
(900)	HC	85/85	75.1±5.7	85/85	75.9±5.9	90/54	75.4±5.2
ADNI2	AD	61/40	75.0±8.5	90/57	75.1±7.8	45/58	74.5±7.3
(844)	HC	64/90	74.3±5.9	92/88	74.0±6.4	68/91	75.6±6.4
AIBL	AD	-	-	28/45	73.6±8.0	-	-
(300)	HC	-	-	107/120	72.9±6.6	-	-
PPMI	PD	82/37	61.6±9.7	78/46	63.0±9.8	68/37	61.6±9.9
(520)	HC	17/17	59.6±13.3	71/34	59.6±10.5	20/13	59.7±11.2
ABIDE	ASD	83/15	12.8±2.6	280/40	16.8±8.2	79/7	18.6±9.7
(1060)	HC	91/27	13.9±3.6	275/55	17.1±7.8	94/14	17.6±8.4
CALSNIC1	ALS	21/25	57.0±11.4	43/28	59.6±10.8	17/1	58.1±9.0
(281)	HC	23/33	50.5±11.9	38/28	57.2±8.1	6/18	53.1±8.4
CALSNIC2	ALS	14/4	54.0±11.8	124/65	60.1±10.2	29/19	62.4±8.2
(545)	HC	18/13	60.1±8.8	120/101	54.9±10.5	10/28	61.7±10.8

3.2.2 Evaluation Metrics

1. t-Distributed Stochastic Neighbor Embedding (t-SNE): It is a nonlinear dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space while preserving the local and global structure of the data [35]. It models each high-dimensional data point as a probability distribution in the lower-dimensional space and minimizes the divergence between the probability distributions. In my case, the t-SNE method takes the input of the proposed 22 features and converts them to two-dimensional space for each MRI scan. The `sklearn.manifold` Python

Table 3.3: Scanning protocol details of the ADNI1, ADNI2, AIBL, PPMI, ABIDE, CALSNIC1, and CALSNIC2 datasets. [FS = Field Strength]

Dataset	Scanning Protocol	MRI Scanner Manufacturer		
		GE	**Siemens**	**Philips**
ADNI1	Model	Genesis Signa, Signa Excite, Signa HDx	Symphony, Sonata, TrioTim, Trio, Avanto, Allegra	Intera Achieva, Intera, Achieva, Gyroscan Intera
	FS	1.5 T / 3.0 T	1.5 T / 3.0 T	1.5 T / 3.0 T
	Flip Angle	8 °	8 °/ 9 °	8 °
	Resolution (mm^3)	$1.0 \times 1.0 \times 1.2$ / $0.94 \times 0.94 \times 1.2$	$1.0 \times 1.0 \times 1.2$ / $1.25 \times 1.25 \times 1.2$	$1.0 \times 1.0 \times 1.2$ / $0.94 \times 0.94 \times 1.2$
ADNI2	Model	Signa HDxt, Signa HDx, Signa Excite, Discovery MR750	Symphony, Skyra, Verio, Avanto, TrioTim	Achieva dStream, Intera, Achieva, Ingenia, Ingenuity
	FS	3.0 T	3.0 T	3.0 T
	Flip Angle	11 °	9 °	9 °
	Resolution	$1.05 \times 1.05 \times 1.2$	$1.05 \times 1.05 \times 1.2$	$1.05 \times 1.05 \times 1.2$
AIBL	Model	-	Avanto, TrioTim, Verio	-
	FS	-	1.5 T / 3.0 T	-
	Flip Angle	-	9 °	-
	Resolution	-	$1.0 \times 1.0 \times 1.2$	-
PPMI	Model	Signa HDxt, Genesis Signa, Signa Architect, Signa Excite, Discovery MR750,	Symphony, Skyra, TrioTim, Prisma, Verio, Espree,	Achieva dStream, Achieva, Intera, Gyroscan NT
	FS	1.5 T / 3.0 T	1.5 T / 3.0 T	1.5 T / 3.0 T
	Flip Angle	8 °/ 11 °/ 13 °/ 15 °	8 °/ 9 °/ 15 °	8 °/ 9 °
	Resolution	$1.0 \times 1.0 \times 1.0$ / $0.94 \times 0.94 \times 1.2$ / $0.94 \times 0.94 \times 0.7$	$1.0 \times 1.0 \times 1.0$ / $1.25 \times 1.25 \times 1.3$ / $0.49 \times 0.49 \times 2.0$	$1.0 \times 1.0 \times 1.0$ / $0.94 \times 0.94 \times 1.2$ / $1.0 \times 1.0 \times 1.2$
ABIDE	Model	Signa Discovery MR750	Allegra, Verio, TrioTim, Prisma,	Achieva, Intera
	FS	3.0 T	3.0 T	3.0 T
	Flip Angle	8 °/ 15 °	7 °/ 8 °/ 9 °/ 10 °	7 °/ 8 °
	Resolution	$1.0 \times 1.0 \times 1.0$ / $0.86 \times 0.86 \times 1.5$ / $1.02 \times 1.02 \times 1.2$	$1.0 \times 1.0 \times 1.0$ / $1.0 \times 1.0 \times 1.33$ / $0.5 \times 0.5 \times 1.2$	$1.0 \times 1.0 \times 1.0$ / $0.98 \times 0.98 \times 1.2$ /
CALSNIC1	Model	Discovery MR750	Prisma, TrioTim	Intera
	FS	3.0 T	3.0 T	3.0 T
	Flip Angle	11 °	8 °	9 °
	Resolution	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$
CALSNIC2	Model	Discovery MR750	Prisma, TrioTim	Achieva
	FS	3.0 T	3.0 T	3.0 T
	Flip Angle	16 °	10 °	10 °
	Resolution	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$	$1.0 \times 1.0 \times 1.0$

library is used to implement t-SNE with the default settings `(n_components = 2, perplexity = 30)`.

2. Uniform Manifold Approximation and Projection (UMAP): It is another non-linear dimension reduction algorithm which focuses on retaining the local structure of the data [36]. In order to map the high-dimensional data to a lower-dimensional space, UMAP creates a topological representation of the data while maintaining the neighbourhood relationships between the data points. The UMAP is implemented with the Python `umap` package with default hyper-parameters `(n_components = 2, n_neighbors= 15, min_dist= 0.1, metric = 'euclidean')`. This study's visual findings are mostly similar for both t-SNE and UMAP. However, the data are more condensed in UMAP and tend to produce more clusters; hence t-SNE is recommended as the first choice.

3. Domain shift distance: The MMD is widely recognized as a prominent metric in DA research to assess the dissimilarities in data distribution between two domains [20]. It can be mathematically defined as the discrepancy between the distributions of domains a and b, $MMD_k^2 = \|\mathbb{E}_p[\phi(x^a)] - \mathbb{E}_q[\phi(x^b)]\|^2_{\mathcal{H}_k}$, where $\mathcal{H}(k)$ represents the Reproducing Kernel Hilbert Space equipped with a kernel function k. A decrease in the MMD distance between the two domains after DA or harmonization process signifies a reduction of domain shift.

4. Domain classification accuracy: Consider a scenario where a random number of samples are selected from two distinct domains, each labelled with its corresponding domain. To evaluate the presence of domain shift or dissimilarity, a domain discriminator or classifier is applied to all the samples, aiming to identify the domain from which each sample originates. The classification outcome serves as a measure of domain shift. A high accuracy in classifying the samples based on their domains implies that the two domains exhibit significant differences, indicating a substantial domain shift. It also supports the robustness of the features used to train the classifier. Conversely, the reduction in domain classification accuracy indicates a decrease

in domain shift, making it more challenging to distinguish between them. Support Vector Machine (SVM) with a linear kernel and Random Forest (RF) with a 500 estimator size are used in the experiments as domain classifiers. These classifiers follow the implementation of the Python `sklearn` package with a five-fold cross-validation setup.

3.2.3 Domain Shift in T1-weighted MRI Data

The multi-center datasets used in this study encompass various factors contributing to domain shift, including scanner manufacturer, model, field strength, image acquisition orientation, resolution, and flip angle. However, when applying the DSMRI framework to these datasets, the resulting clusters primarily demonstrated separation based on the scanner manufacturer parameter, corroborating findings from previous studies [45, 46]. Figure 3.2 presents the visualization of the datasets, considering three distinct domains representing different scanner vendors (e.g., GE, Philips, and Siemens). The first row of Fig. 3.2 depicts the t-SNE plots of the CALSNIC1, CALSNIC2, and ADNI2 datasets, clearly showcasing the separation among data from different manufacturers. In the second row, which pertains to the more challenging ADNI1, PPMI, and ABIDE datasets, some minor overlapping is observed among domains. It might be because of strong similarities in imaging characteristics among those samples. Furthermore, distinct clusters emerge within the same vendor, highlighting the influence of other parameters primarily attributable to the scanner model.

These visual findings are supported by the domain shift distance calculated by MMD, as presented in Table 3.4. Additionally, the domain classification accuracy is consistently around 100% for most cases, which signifies two crucial aspects. Firstly, it highlights the substantial level of domain shift present among the data from different manufacturers, and secondly, it demonstrates the robustness of the features employed in classifying these domains.

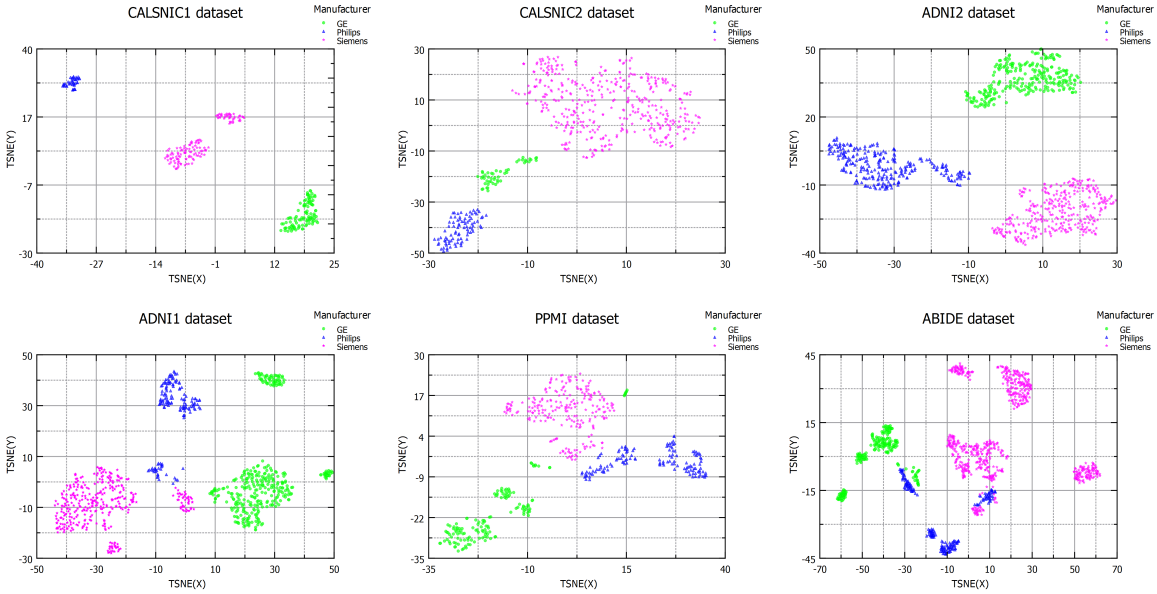


Figure 3.2: t-SNE plots illustrating data distributions across various datasets: CALSNIC1, CALSNIC2, ADNI2, ADNI1, PPMI, and ABIDE. Each data point in the graph corresponds to an individual MRI scan, using three distinct colors to distinguish scans acquired from different scanner manufacturers.

Table 3.4: Domain shift distance in terms of MMD and domain classification accuracy for the ADNI1, ADNI2, PPMI, ABIDE, CALSNIC1, and CALSNIC2 datasets.

Dataset	**Domain shift distance**			**Domain classification accuracy**
	GE vs. Siemens	**GE vs. Philips**	**Philips vs. Siemens**	**GE vs. Siemens vs. Philips**
ADNI1	2.03	0.99	3.01	SVM = 0.99 RF = 1.00
ADNI2	18.06	4.31	7.72	SVM = 0.95 RF = 1.00
CALSNIC1	31.60	369.34	105.59	SVM = 0.99 RF = 1.00
CALSNIC2	3.79	2.23	9.97	SVM = 0.99 RF = 0.99
PPMI	1.35	2.02	1.19	SVM = 0.91 RF = 0.98
ABIDE	2.68	1.78	2.30	SVM = 0.93 RF = 0.99

3.2.4 Effects of Scanner Model

This analysis investigates the impact of different scanner models originating from the same manufacturer. A subset of the ADNI1 dataset comprising five Siemens scanner models, namely Trio, Allegra, Avanto, Sonata, and Symphony, is evaluated to understand the effects of different scanner models. The t-SNE plot in the left panel of Fig. 3.3 illustrates that the data from Avanto, Sonata, and Symphony exhibit similarities in their feature space, indicating comparable imaging characteristics. Additionally, it is worth noting that the Trio and Allegra scanners have a magnetic field strength of 3.0 T, while the other three scanners maintain a field strength of 1.5 T. Moving to the AIBL dataset, it consists of data from three different Siemens scanner models: Avanto, TrioTim, and Verio. The middle panel of Fig. 3.3 shows the t-SNE plot for the AIBL dataset, where data clusters closely align with their respective scanner models. Moreover, the right panel of the diagram further confirms the influence of the magnetic field strength, as the data with a field strength of 1.5 T are separated from the data with a field strength of 3.0 T in the t-SNE map. Lastly, Table 3.5 provides information on the domain shift distance and classification accuracy among the different scanner models, offering insights into the variations between these models.

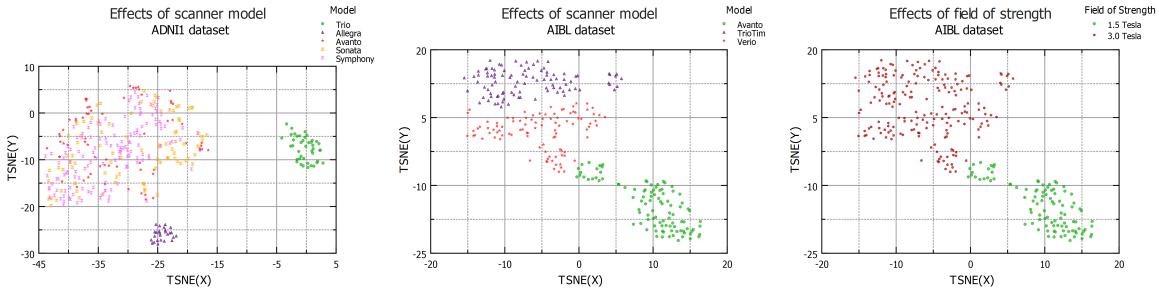


Figure 3.3: t-SNE plots illustrating the domain shift effects resulting from different scanner models of the same manufacturer, observed in the ADNI1 and AIBL datasets.

Table 3.5: Domain shift distance and domain classification accuracy for the ADNI1 (Model 1= Allegra, Model 2= Trio, Model 3= Symphony+Avanto+Sonata), and AIBL (Model 1= Avanto, Model 2= TrioTim, Model 3= Verio) datasets to show the effects of various scanner models.

Dataset	**Domain shift distance**			**Domain classification accuracy**
	Model 1 vs. Model 2	**Model 1 vs. Model 3**	**Model 2 vs. Model 3**	**Model 1 vs. Model 2 vs. Model 3**
ADNI1	4.82	1.80	6.82	SVM = 0.99 RF = 1.00
AIBL	5.02	2.62	0.92	SVM = 0.97 RF = 0.98

3.2.5 Effects of T2-weighted and FLAIR Images

This experiment validates the proposed framework's effectiveness when applied to T2-weighted and FLAIR images. Within the CALSNIC2 dataset, both FLAIR and T2-weighted images were available for the same population. T2-weighted images offer excellent contrast for evaluating pathologies like inflammation, edema, and fluid-filled structures. On the other hand, FLAIR imaging, a variation of T2-weighted imaging, nullifies the signal from fluids like cerebrospinal fluid (CSF) and enhances the visibility of lesions, particularly those adjacent to CSF-filled spaces. Figure 3.4 showcases the t-SNE and UMAP plots for the data derived from these two MRI modalities. Interestingly, the clusters representing different manufacturers are even more distinct for these two modalities compared to T1-weighted images. Table 3.6, presenting the domain shift distance and high domain classification accuracy, provides robust evidence supporting the existence of domain shift in the T2-weighted and FLAIR data while demonstrating the effectiveness of the proposed features.

3.2.6 Effects of Processed Data

In this experiment, the objective is to evaluate the performance of the data after applying commonly used preprocessing neuroimaging pipelines to the CALSNIC1 and CALSNIC2 datasets. As a crucial step in the preprocessing pipeline, I first utilize the FreeSurfer [82] program for skull stripping. Subsequently, the FSL software [132] is

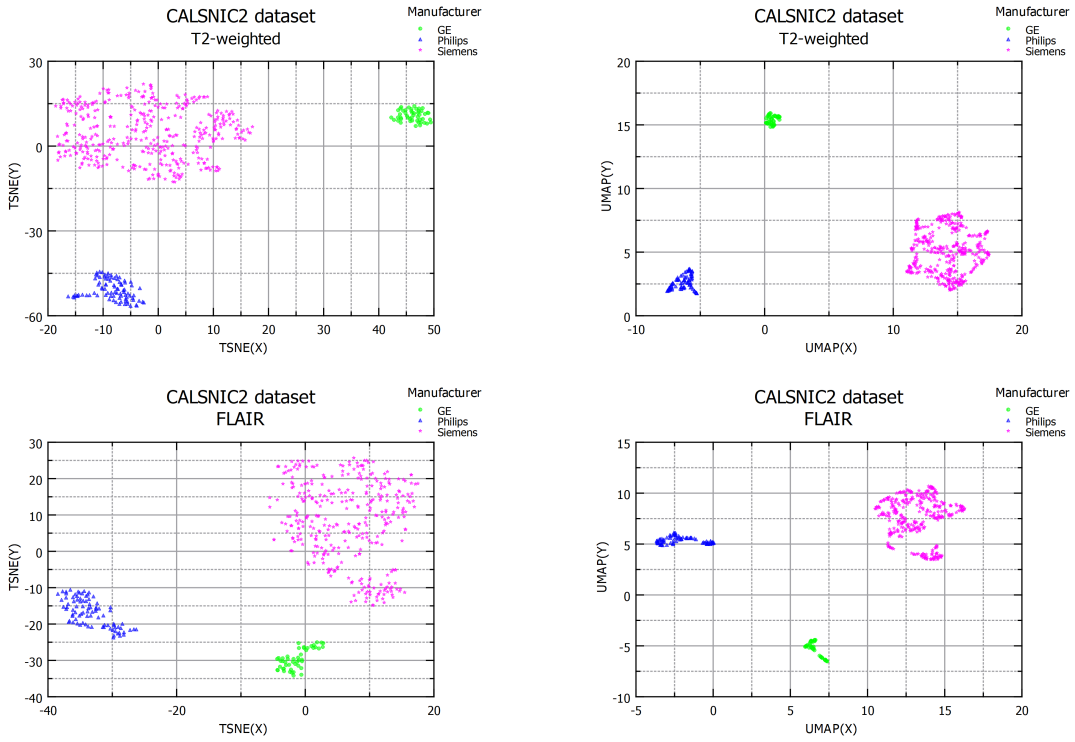


Figure 3.4: t-SNE and UMAP plots illustrating the domain shift effects observed within the CALSNIC2 dataset due to the utilization of T2-weighted and FLAIR images.

Table 3.6: Domain shift distance in terms of MMD and domain classification accuracy for the CALSNIC2 dataset showing the effects of using T2-weighted and FLAIR images.

Dataset	**Domain shift distance**			**Domain classification accuracy**
	GE vs. Siemens	**GE vs. Philips**	**Philips vs. Siemens**	**GE vs. Siemens vs. Philips**
CALSNIC2 T2-weighted	143.75	203.98	130.39	SVM = 1.00 RF = 1.00
CALSNIC2 FLAIR	9.57	6.08	41.73	SVM = 0.98 RF = 0.99

employed to register the MRI scans to the MNI-152 space, ensuring the standardized image and voxel dimensions across all scans. Following these preprocessing steps, I generate t-SNE diagrams to visualize the processed data, as depicted in Fig. 3.5. The visualizations reveal that domain shift remains prevalent in the dataset despite the application of preprocessing techniques.

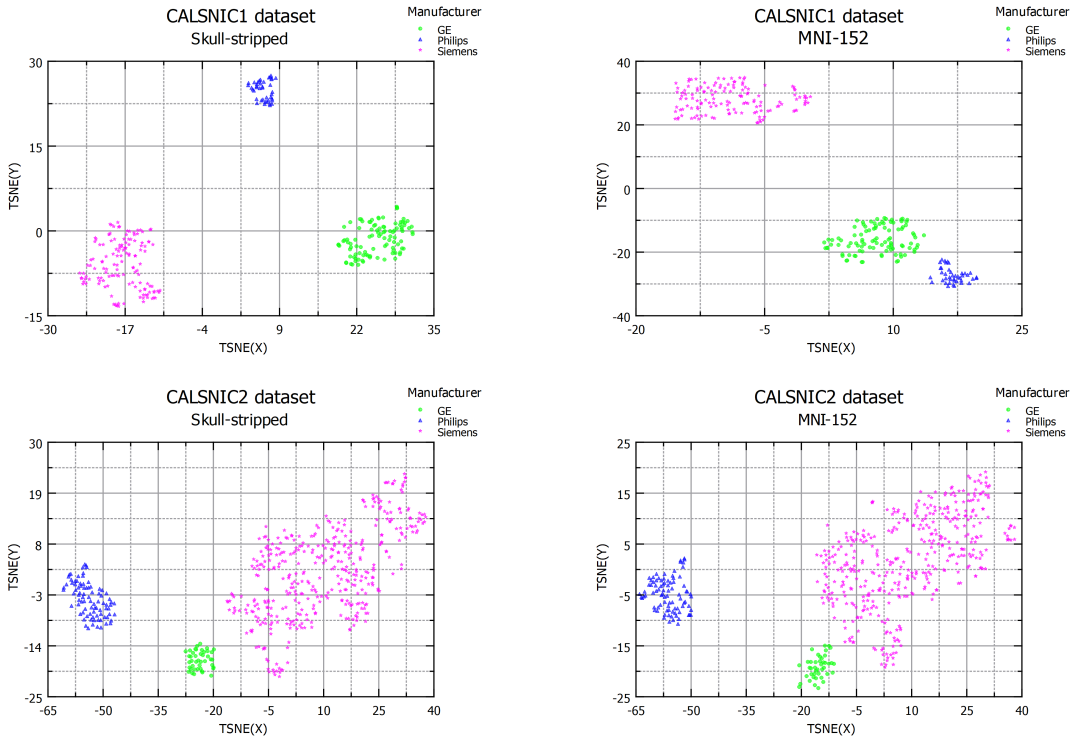


Figure 3.5: t-SNE plots for the CALSNIC1 and CALSNIC2 datasets showing the effects of data after performing skull stripping and registration to MNI-152 template.

To further confirm the presence of domain shift, Table 3.7 presents the domain shift distance between pairs of domains, along with a domain classification accuracy of nearly 100%. These findings provide evidence of the substantial impact of domain shift within the dataset, emphasizing the robustness of the proposed features, which consistently demonstrate their efficacy even with the processed data.

3.2.7 Feature Importance

This section examines the significance of different proposed features across various datasets and data types. To accomplish this, I employ an RF classifier and extract the feature importance ranking from the model. The ranking of the features is presented in Fig 3.6, where the upper left panel displays the average scores of six large datasets utilized in the study. Similarly, the upper right panel depicts the results obtained from the average scores of the processed data from the CALSNIC2 dataset. Interestingly,

Table 3.7: Domain shift distance in terms of MMD and domain classification accuracy for the CALSNIC1 and CALSNIC2 datasets showing the effects of data after performing skull stripping and registration to MNI-152 template.

Dataset	Domain shift distance			Domain classification accuracy
	GE vs. Siemens	GE vs. Philips	Philips vs. Siemens	GE vs. Siemens vs. Philips
CALSNIC1 Skull-stripped	37.86	13.29	150.25	SVM = 1.00 RF = 1.00
CALSNIC1 MNI-152	53.97	3.54	250.46	SVM = 1.00 RF = 1.00
CALSNIC2 Skull-stripped	7.88	5.90	39.92	SVM = 0.99 RF = 0.99
CALSNIC2 MNI-152	4.16	6.21	77.24	SVM = 0.98 RF = 0.98

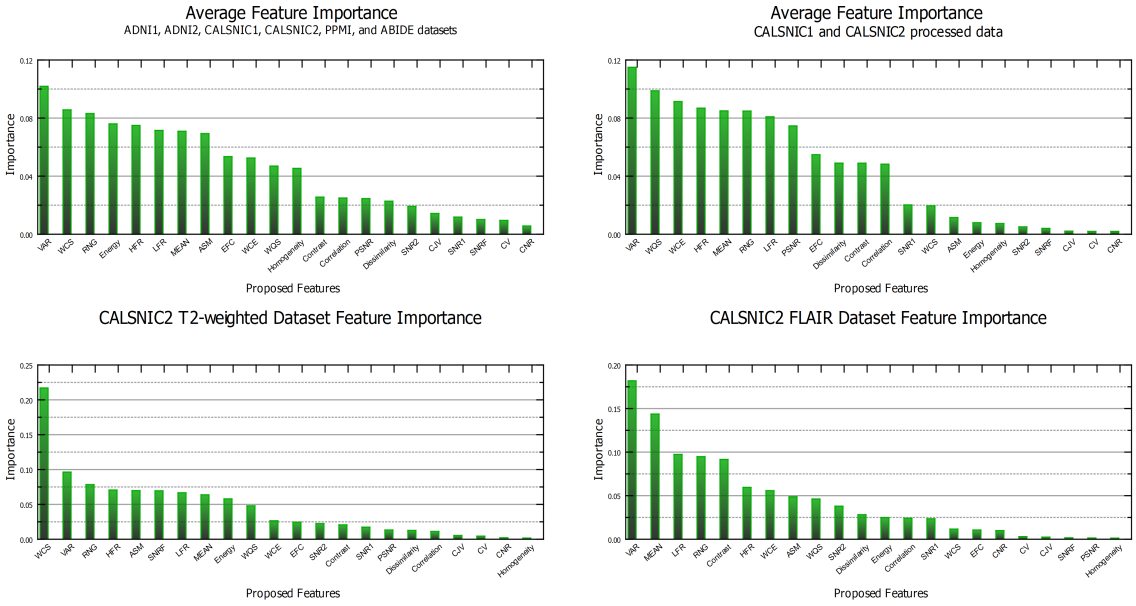


Figure 3.6: Feature importance ranking across various datasets and data types, assessing domain shift presence through prioritizing the 22 proposed features.

the 'VAR' feature consistently achieves the highest ranking in both cases. The frequency domain features, namely 'HFR' and 'LFR,' demonstrate notable importance, while the spatial domain features, such as 'RNG,' 'MEAN,' and 'EFC,' also exhibit promising significance. The wavelet and texture domain features mostly occupy the middle area of the ranking chart. Furthermore, the bottom left and right panels illustrate the outcomes obtained from the CALSNIC2 T2-weighted and FLAIR image

datasets, respectively. In both cases, features such as 'VAR,' 'RNG,' 'MEAN,' 'HFR,' 'LFR,' 'ASM,' and 'WQS' secure positions in the top 10 of the ranking, emphasizing their consistent importance across different data types.

3.2.8 Comparison

The comparative evaluation of the proposed DSMRI framework involves two related methods, namely DomainATM and MRQy, with a focus on visualizing the data using t-SNE plots. Figure 3.7 illustrates the comparison results for three large-scale challenging datasets (e.g., ADNI1, PPMI, and ABIDE). The first column displays the outcomes obtained from DomainATM, revealing inferior performance in clustering the three dominant domains. This can be attributed to the fact that the features

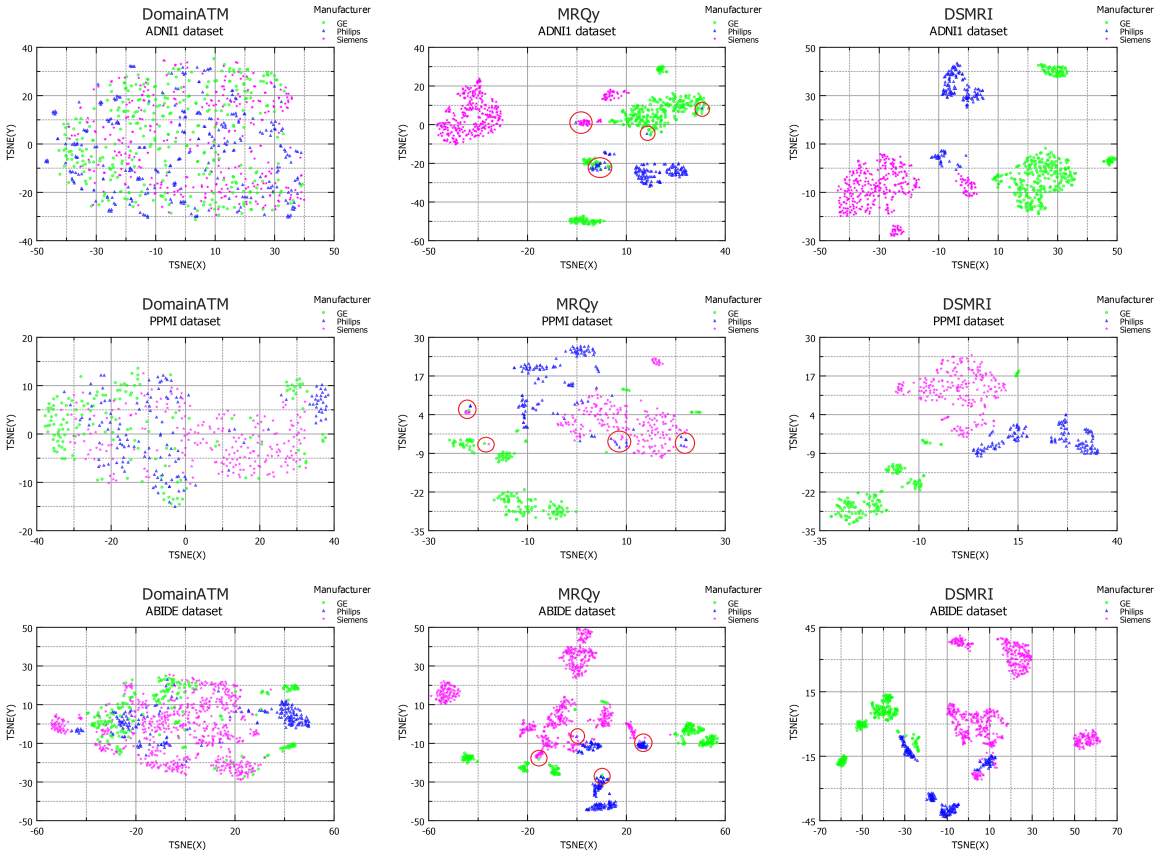


Figure 3.7: Comparison of the proposed framework with two prior approaches visualizing data distribution through t-SNE plots for the challenging ADNI1, PPMI, and ABIDE datasets.

utilized by DomainATM, which are the grey matter volumes of different ROIs, do not exhibit a strong correlation with domain shift measurement. Moving to the middle column, the t-SNE diagram generated by MRQy demonstrates a significant improvement in grouping the data based on scanner vendor. However, upon closer observation (shown in red circles), it becomes apparent that a noticeable amount of data either adopts unexpected positions or slightly deviates from the main clusters, suggesting the presence of weaknesses in their features. Finally, in the last column, the proposed DSMRI approach demonstrates a significant superiority over both DomainATM and MRQy in accurately clustering data from different manufacturers or domains. This compelling performance highlights the strength of the features introduced by DSMRI, which exhibit strong correlations with quantifying the degree of domain shift.

3.3 Summary

The proposed DSMRI framework, explicitly designed to analyze the presence of domain shift in multi-center MRI datasets, to my best knowledge, offers several significant contributions for the first time. Firstly, DSMRI integrates insights from diverse domains, including spatial, frequency, wavelet, and texture analysis. This multi-domain approach fortifies the framework's ability to capture various aspects of domain shift. Secondly, deriving the features from the frequency domain to capture low and high-frequency image information and incorporating wavelet domain features to measure sparsity and energy within wavelet coefficients enhance the robustness of domain shift analysis. Thirdly, using visualization techniques such as t-SNE and UMAP enriches the framework's ability to visually represent and interpret domain shift effects. Fourthly, estimating domain shift distance, domain classification accuracy, and the ranking of significant features add a rigorous quantitative evaluation of domain shift. Lastly, the efficacy of DSMRI is validated through extensive experimental evaluations conducted on seven large-scale multi-site neuroimaging datasets. This real-world validation showcases the practical applicability of the proposed framework.

In the field of neuroscience research, multi-center neuroimaging studies require robust, efficient, and reliable techniques to address the non-biological sources of data variation. ML-based approaches often yield inconsistent results when dealing with data acquired from different MRI scanner models and scanning protocols. This study makes a significant contribution by presenting a simple yet effective unsupervised framework for quantifying the degree of domain shift. After examining a wide range of large multi-center MRI datasets, this study explores the impacts of different scanner manufacturers, models, field strengths, and resolutions in the context of domain shift. Furthermore, the proposed framework demonstrates its adeptness in identifying domain shift, not only in preprocessed T1-weighted MRI data but also across T2-weighted and FLAIR modalities. The findings of this study have important implications for advancing the field of medical imaging and enabling more reliable analysis of multi-center MRI datasets. Moreover, DA and harmonization methods can utilize the proposed framework to validate the effectiveness of their approaches in reducing or eliminating domain shift.

Chapter 4

Deep Domain Shift Analyzer for MRI

The proposed DSMRI method described in the previous chapter was designed for structural MRI. Therefore, it could not exhibit robustness in identifying domain shift for some advanced MRI modalities such as functional MRI (fMRI) and diffusion-weighted imaging (DWI). To address this limitation, I propose another novel framework called *Deep Domain Shift analyzer for MRI* (DeepDSMRI), designed explicitly to comprehend the extent of domain shift in multi-modal MRI datasets [133]. Utilizing pre-trained deep models as feature extractors, DeepDSMRI provides adequate insights into the existence of domain shift for diverse MRI modalities, including structural, functional, and diffusion-weighted images. The datasets and evaluation metrics are the same as previous chapter. The source code has been made publicly available at https://github.com/rkushol/DeepDSMRI.

4.1 Proposed Method

4.1.1 Overview

The overall workflow of the proposed DeepDSMRI framework is depicted in Fig. 4.1. This methodology is driven by the concept of harnessing the knowledge acquired by a deep neural network through pre-training on the extensive ImageNet [37] computer vision dataset and extending its applicability to the medical imaging domain. In the early layers of the deep network, fundamental low-level features such as edges, colors,

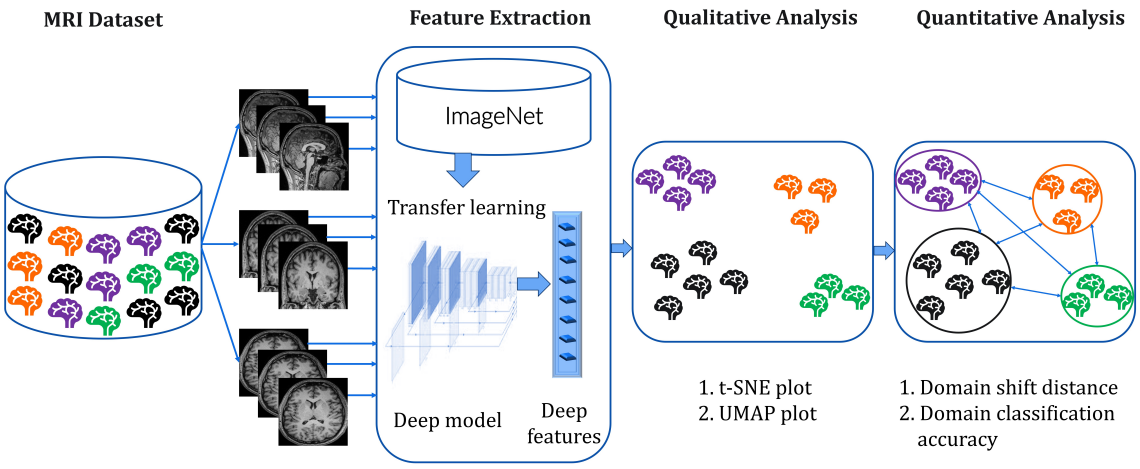


Figure 4.1: The assorted colors of the brain icon in the MRI dataset denote the source of the originating data using distinct acquisition protocols. The pre-trained deep models are used as feature extractors, which transfer knowledge from the ImageNet dataset and extract deep features from 2D MRI slices of each sample. The t-SNE and UMAP algorithms display data similarity in a 2D graph using these feature maps. Finally, the domain shift distance using maximum mean discrepancy (MMD) and the classification accuracy of various domains are calculated in quantitative analysis.

and textures are captured. These basic features effectively represent essential patterns inherent in MR images. Moving deeper into the network, mid-level features emerge, encapsulating more intricate patterns or textures formed by combinations of low-level features, such as certain shapes or object segments. The deeper layers of the network are dedicated to capturing high-level semantic features, offering representations of complex structures and object components. These layers excel in grasping the broader context, spatial relationships, and semantic intricacies of objects within the images. Leveraging pre-trained deep models as feature extractors facilitates the extraction of basic image patterns and representations without requiring the fine-tuning of new data. An example of an output MR image generated through different layers of a deep model is illustrated in Fig. 4.2. My proposed approach excludes the final output layer responsible for task-specific predictions. This strategic choice enables the framework to comprehend the extent of domain shift in MRI data, providing valuable insights into the variations across different imaging domains. The following section details the deep models employed as feature extractors. I apply nine widely used deep networks

using PyTorch *timm* library [134], and based on the qualitative results, most of them are capable of visualizing domain shift in MRI data.

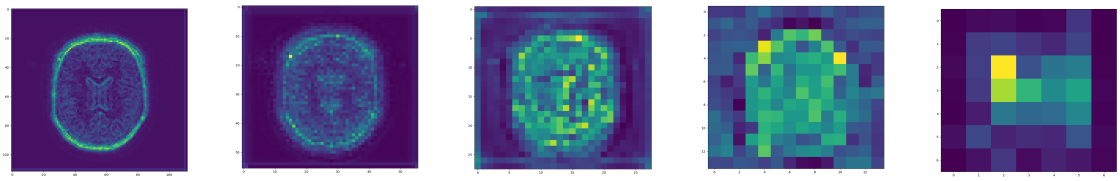


Figure 4.2: Visualization of feature representation for a 2D MRI slice using various layers of the pre-trained ResNet50 deep model. The images from left to right display the output of feature representation for layers one to five, respectively.

4.1.2 Deep Networks

The analysis encompasses two most popular types of deep networks: Convolutional Neural Networks (CNNs) and transformer-based architectures. First, ResNet [135] is a widely used CNN architecture that introduces residual connections to improve training speed. Out of several versions of ResNet, I use ResNet50, which generates 2048 deep features for each image. Next, DenseNet [136] is well-known for its densely connected blocks, where each layer takes input from all prior layers in the block. This dense connectivity encourages feature reuse with parameter efficiency which alleviates the vanishing gradient problem. I employ DenseNet169, which yields 1664 deep features. Another deep model, MobileNetV2 [137], is designed for devices with limited computational resources, emphasizing efficiency and speed using inverted residuals and linear bottlenecks. Lastly, EfficientNet [138] employs a compound scaling strategy to optimize the model's efficiency by balancing trade-offs between model size and performance. The final size of features produced by EfficientNet and MobileNetV2 are 1792 and 1280, respectively.

The Vision Transformer (ViT) [38] has revolutionized image classification by treating images as sequences of patches, transforming the task into a sequence-to-sequence problem. It employs self-attention (SA) mechanism for capturing global and local dependencies within the image. In the implementation, the *vit_base_patch*16_224

version is employed which generates 768 deep features. Building on ViT's success, the Swin transformer [86] introduces a hierarchical design with shifted windows to effectively capture information across different scales, enhancing the model's performance on both global and local features. I prefer the *swin_base_patch*4_*window*7_224 version, which yields 1024 features. Another transformer-based model, DeiT [88] focuses on training models with limited data by leveraging knowledge distillation. It utilizes a teacher-student framework, where a larger transformer (teacher) transfers knowledge to a smaller transformer (student). The implementation follows the *deit_base_patch*16_224 version, which creates 768 features. Some models combine the strengths of both CNNs and transformers. The CoaT [139] model, for instance, enhances the multi-scale and contextual modeling capabilities of ViT by introducing a co-scale mechanism and convolutional attention module. I employ the *coat_mini* version with 216 deep features. Lastly, ConViT [140] introduces gated positional self-attention (GPSA), another hybrid approach of combining CNNs and ViT with 768 features for the *base* configuration. Although the proposed framework employs these deep methods separately, their performance is comparable. However, based on slightly superior qualitative results on challenging datasets such as ABIDE and PPMI, the experimental analysis section primarily showcases the performance of the Swin transformer.

4.2 Experiments

4.2.1 Domain Shift in T1-weighted MRI Data

The MRI data employed in this study encompass several aspects that contribute to domain shift, as illustrated in Table 3.3. Nevertheless, upon analyzing the DeepDSMRI framework across the datasets of T1-weighted images, the resulting clusters predominantly exhibit partitions based on the scanner manufacturers. This observation aligns with conclusions from prior research studies [43, 45, 46]. Three scanner vendors are

treated as distinct domains in the visualization of four datasets, as shown in Fig. 4.3. The t-SNE plots for the remaining datasets are presented in the comparison section. Notably, for the CALSNIC1 dataset, a subcluster within the Siemens data reflects significant variations in scanner models (*i.e.*, Prisma and TIM Trio). Similarly, for the AIBL dataset, different scanner models of Siemens show uniqueness, and the data can be separated based on the field strength (*i.e.*, 1.5T and 3.0T) as well. The visual findings align with the MMD-calculated domain shift distance, which is shown in Table 4.1. Furthermore, in the majority of cases, the domain classification accuracy continuously hovers around 100%. This high accuracy reinforces two important points: firstly, it highlights a significant amount of domain shift between scanner manufacturers, and secondly, it emphasizes the effectiveness of the deep features used to detect these domains.

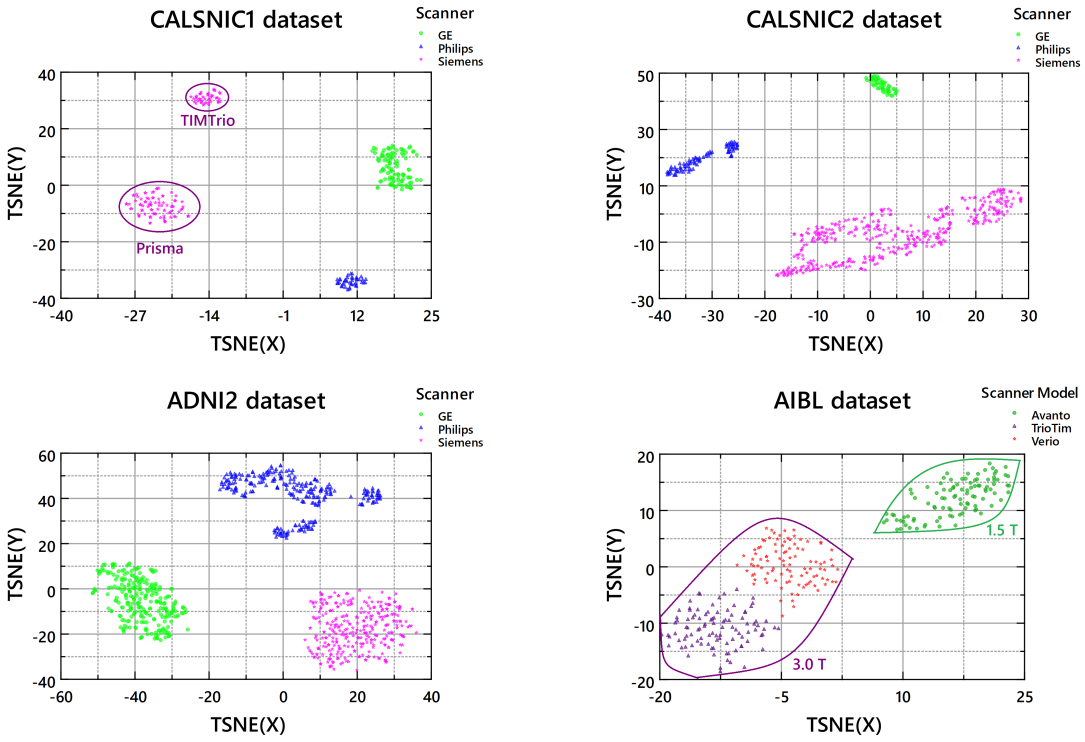


Figure 4.3: t-SNE graphs illustrating domain shift of T1-weighted MRI data for the CALSNIC1, CALSNIC2, ADNI2, and AIBL datasets. Different colors represent data originating from three scanner manufacturers, except AIBL, where colors denote data from different models of the Siemens scanner.

Table 4.1: Domain shift distance for the ADNI1, ADNI2, CALSNIC1, CALSNIC2, PPMI and ABIDE datasets in terms of MMD and domain classification accuracy.

Dataset	Domain shift distance			Domain classification accuracy
	GE vs. Siemens	Philips vs. GE	Siemens vs. Philips	GE vs. Philips vs. Siemens
ADNI1	31.21	57.20	73.37	SVM = 0.99, RF = 0.97
ADNI2	307.14	149.84	172.02	SVM = 1.00, RF = 1.00
PPMI	226.42	236.56	102.96	SVM = 0.98, RF = 0.97
ABIDE	10.62	9.01	29.57	SVM = 0.99, RF = 0.99
CALSNIC1	342.66	255.26	848.75	SVM = 1.00, RF = 1.00
CALSNIC2	431.22	324.95	224.25	SVM = 1.00, RF = 1.00

4.2.2 T2-weighted, FLAIR, fMRI and DWI Data

This experiment substantiates the efficacy of the proposed technique when applied to diverse MRI modalities, including FLAIR, DWI, fMRI, and T2-weighted images. The CALSNIC2 dataset facilitates MR images of these modalities for the same population. Due to exceptional contrast, T2-weighted sequences are valuable for assessing pathologies such as inflammation and edema. Conversely, FLAIR, a variant of the T2-weighted image, suppresses signals from fluids and enhances lesion visibility, especially near CSF-filled spaces. DWI is an advanced MRI modality that measures the random motion of water molecules within tissues and provides information about the microstructural organization of tissues. Lastly, fMRI measures the hemodynamic response, reflecting increased blood flow to active brain regions during performing tasks or resting. Figure 4.4 displays t-SNE plots representing data for the aforementioned MRI modalities. Remarkably, for these modalities, the clusters distinguishing various scanners are more prominent than T1-weighted data. The higher MMD distance and 100% classification accuracy between different domains strongly support the presence of domain shift, as noted in Table 4.2. This evidence further reinforces how well the deep features express the domain shift in MRI data. To our knowledge, no existing framework is capable of assessing the extent of domain shift for fMRI and DWI data. By employing the handcrafted features of the previous work of DSMRI [16] for these

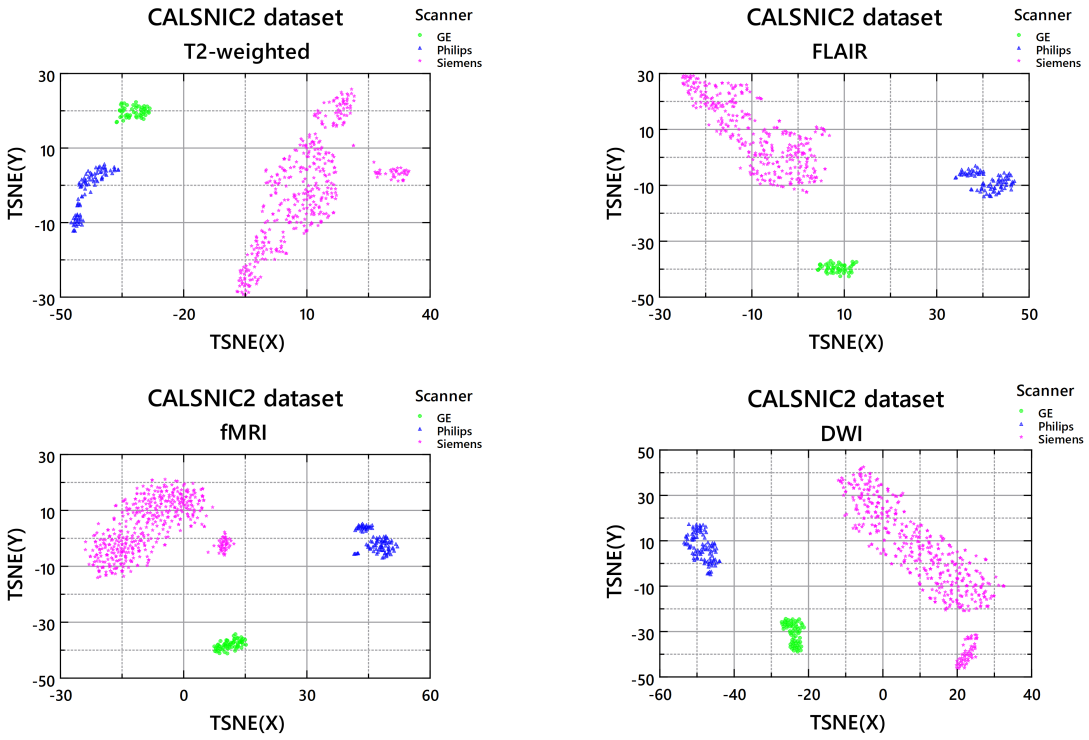


Figure 4.4: t-SNE plots demonstrating domain shift of T2-weighted, FLAIR, fMRI and DWI MR images on the CALSNIC2 dataset using the proposed framework.

two modalities, I provide some failure cases as shown in Fig. 4.5.

Table 4.2: Domain shift distance for the CALSNIC2 dataset demonstrating the impacts of T2-weighted, FLAIR, fMRI, and DWI in terms of MMD and domain classification accuracy.

Modality	**Domain shift distance**			**Domain classification accuracy**
	GE vs. Siemens	**GE vs. Philips**	**Philips vs. Siemens**	**GE vs. Siemens vs. Philips**
T2-weighted	612.08	115.11	283.15	SVM = 1.00, RF = 1.00
FLAIR	68.47	133.22	67.41	SVM = 1.00, RF = 1.00
fMRI	280.79	294.12	328.87	SVM = 1.00, RF = 1.00
DWI	558.65	345.32	756.64	SVM = 1.00, RF = 1.00

4.2.3 Comparison

Comparative analysis involves evaluating the performance of the proposed framework through a comparison with two related methods, DomainATM and DSMRI, using the t-SNE data visualization tool. Figure 4.6 displays the comparison outcomes for three

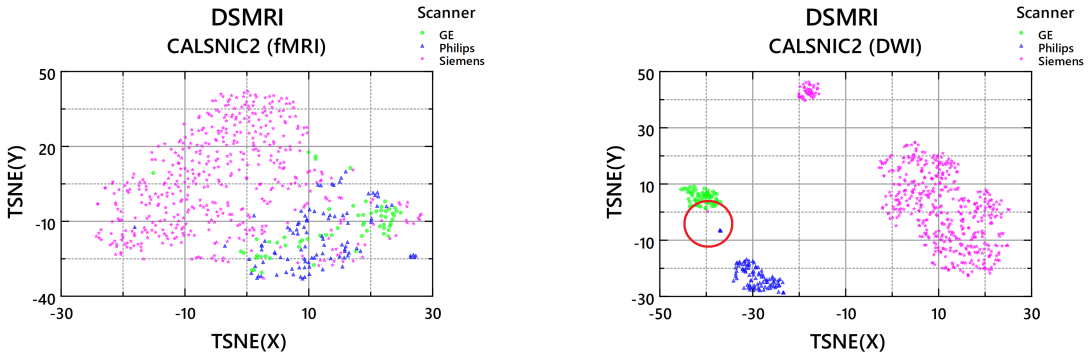


Figure 4.5: On the left side, the t-SNE plot illustrates domain shift for the fMRI data using the features of a prior study called DSMRI, revealing its limitations in identifying domain shift within the fMRI modality. Additionally, the t-SNE plot on the right side highlights minor failures, indicated by red circle, observed in the DWI data on the CALSNIC2 dataset.

challenging datasets (PPMI, ABIDE, and ADNI1). Results from DomainATM are shown in the first column, clearly indicating suboptimal performance in identifying domain shift among different scanners' data. This is probably due to DomainATM's use of grey matter volume features, which correlate weakly to domain shift estimation. Next, the DSMRI method significantly improves grouping data based on scanner vendors. However, closer scrutiny (highlighted in the red circle) reveals that one scanner manufacturer's data overlapped with that of another manufacturer, indicating potential feature deficiencies. In the last column, the proposed DeepDSMRI outperforms DomainATM and DSMRI in precisely grouping data from various scanners.

4.3 Summary

DeepDSMRI offers a simple yet efficient unsupervised framework for analyzing domain shift in MRI data, utilizing different deep models pre-trained with the ImageNet dataset and requiring no training on existing MRI data. The experimental results demonstrate robustness across a spectrum of MRI modalities, including structural (*e.g.*, T1-weighted and T2-weighted), DWI, and fMRI. To our knowledge, this work is the first to analyze and quantify domain shift in multi-modal MRI data using deep

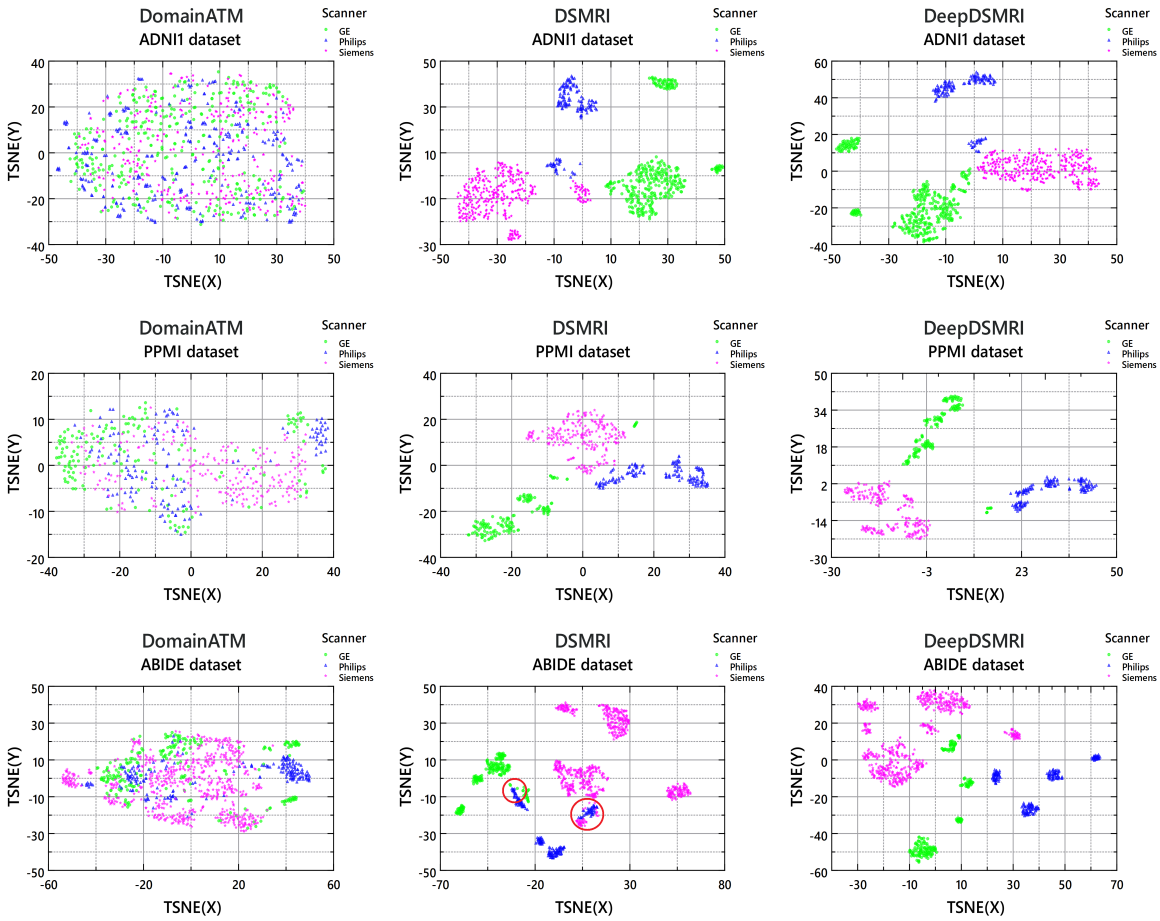


Figure 4.6: Comparative analysis of various methods illustrating domain shift of T1-weighted MRI data across the challenging ABIDE, PPMI, and ADNI1 databases through t-SNE plots.

learning. Moreover, the proposed framework incorporates visualization tools (*e.g.*, t-SNE and UMAP) to illustrate grouping similar data and isolating dissimilar data into distinct clusters. Furthermore, the quantitative analysis encompasses the classification accuracy between domains and the domain shift distance using MMD. The efficacy of the proposed DeepDSMRI is demonstrated through experimental assessments conducted on seven extensive multi-center neuroimaging databases.

This study significantly contributes by introducing a simple yet valuable unsupervised approach to quantify the extent of domain shift in MRI data. More importantly, DeepDSMRI demonstrates its efficacy in determining domain shift not only in structural MRI but also in other advanced MRI modalities such as fMRI and DWI. Last

but not least, the proposed framework can serve as a valuable tool for DA and harmonization methods to verify the effectiveness of their strategies in mitigating domain shift.

Chapter 5

Effects of Scanner Manufacturer

This chapter investigates the performance of multiple disease classification tasks using multi-center MRI data obtained from three widely used scanner manufacturers: GE Healthcare, Philips, and Siemens. I thoroughly examine how variations in MRI scanner manufacturers affect different classification tasks using various deep learning (DL) models. Additionally, this study evaluates whether applying a ComBat-based harmonization technique can improve classification performance after 3D image-level harmonization. Furthermore, I introduce a novel transformer-based classification framework named ADDFormer (Alzheimer's Disease Detection using Transformer). This approach is pioneering in applying a transformer-based deep model for Alzheimer's disease classification, demonstrating superior performance compared to existing popular DL methods. The source code is available at https://github.com/rkushol/ADDFormer.

5.1 Proposed Method

5.1.1 ADDFormer

The proposed ADDFormer method includes a process to select a range of slices, utilizing two transformer networks for selecting features in the spatial and Fourier domains, fusing the gathered features with a third transformer network and performing a majority voting on the predictions to finalize the results. Figure 5.1 shows the architecture of the proposed method.

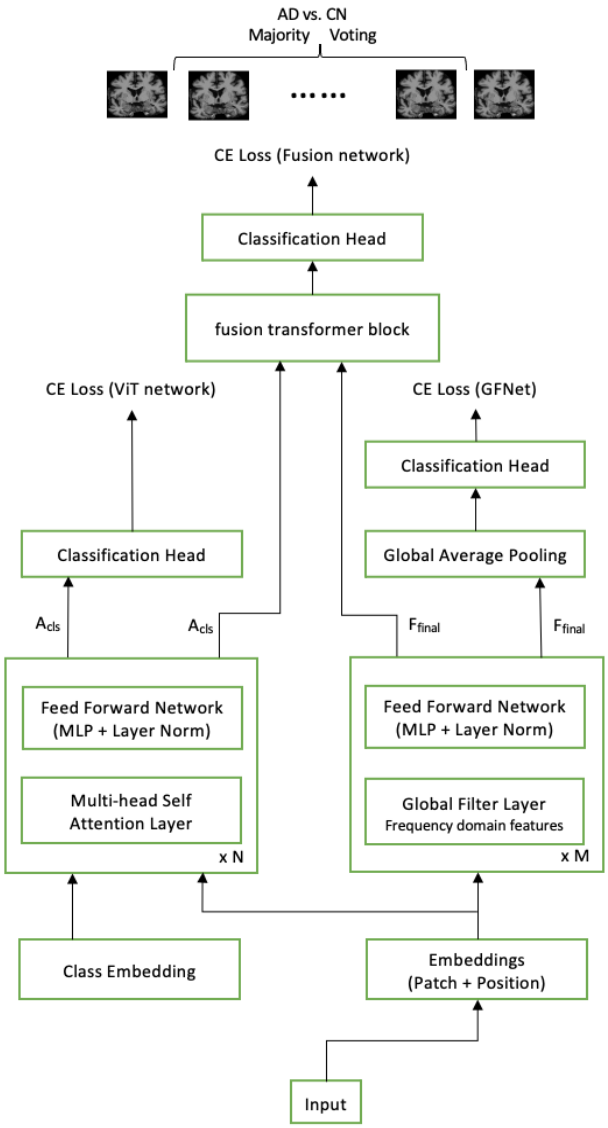


Figure 5.1: The overall workflow of the proposed ADDFormer architecture.

After performing an empirical analysis among the three different axes (coronal, axial and sagittal) of 3D brain MRI, the best performance is achieved by employing the coronal slices. A total of 15 consecutive 2D coronal slices from the central part of the plane are used to train the network.

Vision transformer (ViT) [38] is the first work that applies the transformer [85] to vision tasks. Different from the conventional deep learning-based methods that utilize convolution, which is hard to capture the long-term dependencies, to extract features from images, ViT exploits self-attention to handle such a problem. However, due to

the high computational cost of matrix multiplication in the self-attention, it embeds the image into patch tokens and utilizes patch tokens as input of transformer block to reduce the computation. The attention output of the extra class embedding A_{cls} is passed into the classification head (several fully-connected layers) for final prediction.

MRI data is originally acquired in the frequency domain and then transformed into the spatial domain. Thus, it is more natural to perform feature selection in the frequency domain. However, due to the high resolution, the computational cost of directly applying the network on images in the frequency domain is high. GFNet [87] follows the downsampling strategy of ViT and adopts the fast Fourier transform (FFT) on the embedded patch tokens. Instead of exploiting the output of the extra class embedding as the input of the head, GFNet uses the average pooling of the final feature map F_{final} as the input of the classification head.

To take advantage of information from both the spatial domain and frequency domain, I propose a new fusion transformer block to fuse the A_{cls} and F_{final}, further improving the classification accuracy. As shown in Fig. 5.1, the A_{cls} and the F_{final} are used as the class embedding and features, respectively, of the fusion transformer. The architecture of the fusion transformer block is exactly the same as the transformer block in ViT, except that the class embedding and features are from different networks. Finally, the attention output of the A_{cls} is regarded as the input of the final classification head. Both of the ViT and GFNet are pre-trained on the ImageNet dataset[37]. I first separately train each of the two base transformer networks. Then, the trained weights are used to initialize the combined architecture and train the fusion head.

Lastly, the affected tissue or region is not expected to be present in all the selected slices. In other words, it is not possible to select slices that will always contain distinguishable features without manual effort from an expert. To mitigate the false positive response from those slices with non-significant features, I have taken advantage of majority voting. Therefore, the final classification of a particular subject is

based on the detection of the class that is present in the majority of the slices in the pre-determined range.

5.1.2 Existing DL Models

To assess the performance of various classification tasks across different neuroimaging datasets with distinct scanner vendors, I employ both 2D and 3D DL architectures. Firstly, three widely recognized and successful networks are utilized: ResNet [135], ShuffleNetV2 [141], and MobileNetV2 [137]. The Residual Network (ResNet), a prominent and influential DL model, was introduced by He *et al.* [135]. A pivotal contribution of ResNet is the introduction of "identity shortcut connections," creating alternate pathways for gradient flow and addressing the vanishing gradient problem in deep CNNs. The fundamental building block of MobileNet [137] is depthwise separable convolution, which comprises depthwise convolution and pointwise convolution. Depthwise convolution applies distinct kernels to each input channel, while pointwise convolution employs 1×1 convolution kernels.

ShuffleNet [141], designed to accommodate mobile device computing limitations, relies on pointwise group convolution and channel shuffling to maintain accuracy while significantly reducing computational load. Subsequently, I employ two customized models designed explicitly for AD classification. Qiu *et al.* [4] introduced a 3D customized Fully Convolutional Network (FCN) consisting of six convolutional blocks and then integrated both neuroimaging and clinical data using Multilayer Perceptron (MLP) networks. However, my study only employs their FCN model to handle neuroimaging data. Last but not least, my proposed method, ADDFormer [142] which utilizes frequency and spatial domain features in an innovative manner. Figure 5.2 illustrates the processing pipeline for both 2D and 3D frameworks. In the case of 3D networks, after preprocessing, DL models analyze the entire 3D brain MRI data to extract features for the final class prediction. Conversely, for 2D networks, 15 coronal slices from the central position are assessed for feature extraction. The final

classification decision is determined by the majority voting of class predictions from these coronal slices of a subject, similar to the approach used in ADDFormer.

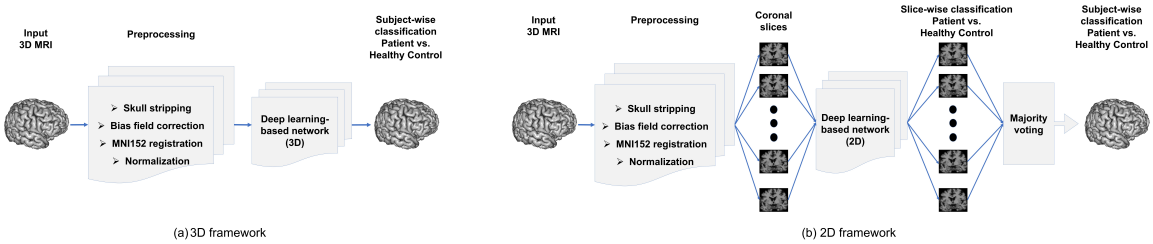


Figure 5.2: The processing pipeline used in the study to carry out different disease classification tasks with different DL networks.

5.2 Experiments

5.2.1 Datasets

The Alzheimer's Disease Neuroimaging Initiative (ADNI)[1] [127] and the Parkinson Progression Marker Initiative (PPMI)[2] [129] represent two prominent and extensively studied publicly available datasets in the field of AD and PD detection, respectively. Additional information is accessible at ADNI and PPMI. I attained approval to use the ADNI and PPMI data in the present study. The Canadian ALS Neuroimaging Consortium (CALSNIC)[3] [131] is the only prospective, multi-center and multimodal longitudinal study of ALS using harmonized clinical and imaging protocols across its sites. The CALSNIC study was conducted with the approval of each participating site's HREB, and informed consent was obtained from the participants. This study leverages T1-weighted MR images, commonly used for standard structural imaging, acquired from three distinct MRI manufacturers (GE, Philips, and Siemens) across the aforementioned datasets. The acquisition orientation of all the MRI data used in our study is sagittal. I employ two versions of ADNI: ADNI1 and ADNI2, consisting of 1638 and 865 MRI scans, respectively. Additionally, my study enlists 528 samples

[1](http://adni.loni.usc.edu/)
[2](http://www.ppmi-info.org/)
[3](https://calsnic.org/)

from PPMI and 545 samples from the CALSNIC2 datasets. CALSNIC1 data were excluded from my experiments due to its comparably limited sample size as well as variations in MRI acquisition orientation. An insightful depiction of the demographic composition of the utilized datasets is presented in Table 5.1. Furthermore, Table 3.3 meticulously outlines the divergent scanning protocols linked to different scanner manufacturers.

Table 5.1: Demographic details of the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets. [CN: cognitively normal]

Dataset	**Group**	**MRI Scanner Manufacturer**					
		GE		**Siemens**		**Philips**	
		Sex (M/F)	**Age (Mean±Std)**	**Sex (M/F)**	**Age (Mean±Std)**	**Sex (M/F)**	**Age (Mean±Std)**
ADNI1	AD	80/80	75.5±7.7	80/80	75.0±7.2	60/49	75.7±7.0
	CN	80/80	75.1±5.7	80/80	75.9±5.9	109/67	75.4±5.2
	MCI	150/100	75.3±7.6	150/100	76.1±7.0	150/63	75.9±7.5
ADNI2	AD	62/41	75.0±8.5	100/57	75.1±7.8	48/58	74.5±7.3
	CN	80/82	74.3±5.9	100/57	74.0±6.4	80/100	75.6±6.4
PPMI	PD	83/40	61.6±9.7	78/46	63.0±9.8	70/37	61.6±9.9
	CN	17/17	59.6±13.3	72/35	59.6±10.5	20/13	59.7±11.2
CALSNIC2	ALS	14/4	54.0±11.8	124/65	60.1±10.2	29/20	62.4±8.2
	CN	18/13	60.1±8.8	120/101	54.9±10.5	12/25	61.7±10.8

5.2.2 Preprocessing

A straightforward, rapid, and commonly employed preprocessing pipeline is implemented to prepare the original 3D T1-weighted brain MRI data for disease classification tasks. The process begins with a standard operation known as skull stripping, aimed at eliminating the unnecessary skull region. This task is achieved using the *FreeSurfer* program [82] (Command: `mri_synthstrip -i input_image -o stripped_image`) [6]. Subsequently, we perform N4 bias field correction using the *SimpleITK* library's `N4BiasFieldCorrectionImageFilter` class to rectify low frequency intensity non-uniformity in the MRI data [143]. The Symmetric normalization (SyN) registration technique, implemented through *ANTsPy* [144], is then employed

to align each scan with MNI-152 standard space, using `lanczosWindowedSinc` interpolation for transformation. Lastly, I apply *WhiteStripe* intensity normalization using the python `intensity-normalization` package [145]. Upon completing the preprocessing of the original images, their dimensions are transformed to $182 \times 218 \times 182$, and the voxel size is converted to $1 \times 1 \times 1\ mm^3$. This preprocessing procedure typically takes around 5 minutes per scan, with computations performed on an eight-core CPU platform utilizing parallel processing.

5.2.3 Implementation

The DL frameworks employed in this analysis are implemented using PyTorch [146] and executed on a server equipped with 4 NVIDIA RTX A6000 GPUs. The coding of 3D CNN models is based on publicly available implementations, accessible at (https://github.com/xmuyzz/3D-CNN-PyTorch). To enhance training robustness, I employ data augmentation methodologies, including random rotations, flipping, and the mixture of Gaussian noise, to prepare a robust training batch. The optimization process employs the Adam optimizer with an initial learning rate of 0.00005 and a decay rate of 10^{-1} after every 100 iterations. For the ADDFormer model, a patch size of 16×16 is used, and the training spans a total of 300 epochs with a batch size of 16. The final accuracy reported in this study represents the average results from five experiments, each employing distinct training, validation, and test data combinations. The data split ratio is maintained at 70% for training, 15% for validation, and 15% for testing in each experimental setup. The training time of the CNN-based procedures takes approximately 5 hours on a single GPU with 48GB of memory. The classification performance is evaluated using standard statistical metrics, specifically Accuracy (Acc) and F1-score. They are characterized in terms of four key values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The *Acc* metric represents the fraction of accurately identified subjects to the total number of samples in a given dataset, defined as $Acc = \frac{TP+TN}{TP+TN+FP+FN}$. The

F1-score harmonically combines precision and recall, and is mathematically measured as $\textit{F1-score} = 2 \times \frac{precision \times recall}{precision + recall}$. The recall is the ability to identify individuals with a specific condition correctly and is computed as $recall = \frac{TP}{TP+FN}$. The precision reflects the number of relevant items and can be expressed as $precision = \frac{TP}{TP+FP}$.

5.2.4 Scanner Manufacturer Effects

This section presents the results of a series of experiments highlighting the distinctive characteristics of different scanner manufacturers. Initially, I employ three 3D DL-based classification networks (ResNet [135], MobileNetV2 [137], ShuffleNetV2 [141]) using T1-weighted MRI data to classify three distinct scanner manufacturers (GE, Philips, and Siemens). These well-established CNN-based networks demonstrate exceptional accuracy in classifying the scanner manufacturers. For the ADNI1, ADNI2, and CALSNIC2 datasets, the average classification accuracy exceeds 98%, while the accuracy for the PPMI database ranges between 93% and 96% across all the aforementioned frameworks. The classification outcomes, presented as confusion matrices derived from the ResNet architecture for different datasets, are depicted in Fig. 5.3. The corresponding confusion matrices for the ShuffleNetV2 and MobileNetV2 models can be found in Fig. A.1 and A.2 of the Appendix.

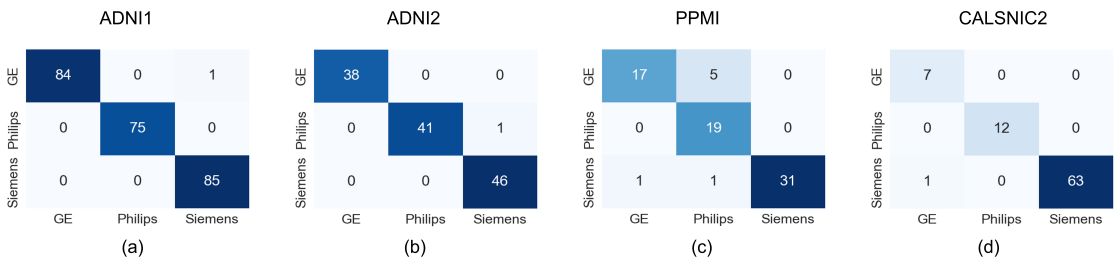


Figure 5.3: MRI scanner manufacturer classification results for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets generated by ResNet model. The classification accuracy is approximately 99% for the (a) ADNI1, (b) ADNI2, and (d) CALSNIC2 datasets whereas the accuracy is around 95% for the (c) PPMI dataset.

Subsequently, I employ t-SNE [35] technique to visualize the data in a 2D space, using features generated by MRQy [57] as presented in Fig. 5.4. The t-SNE is a non-

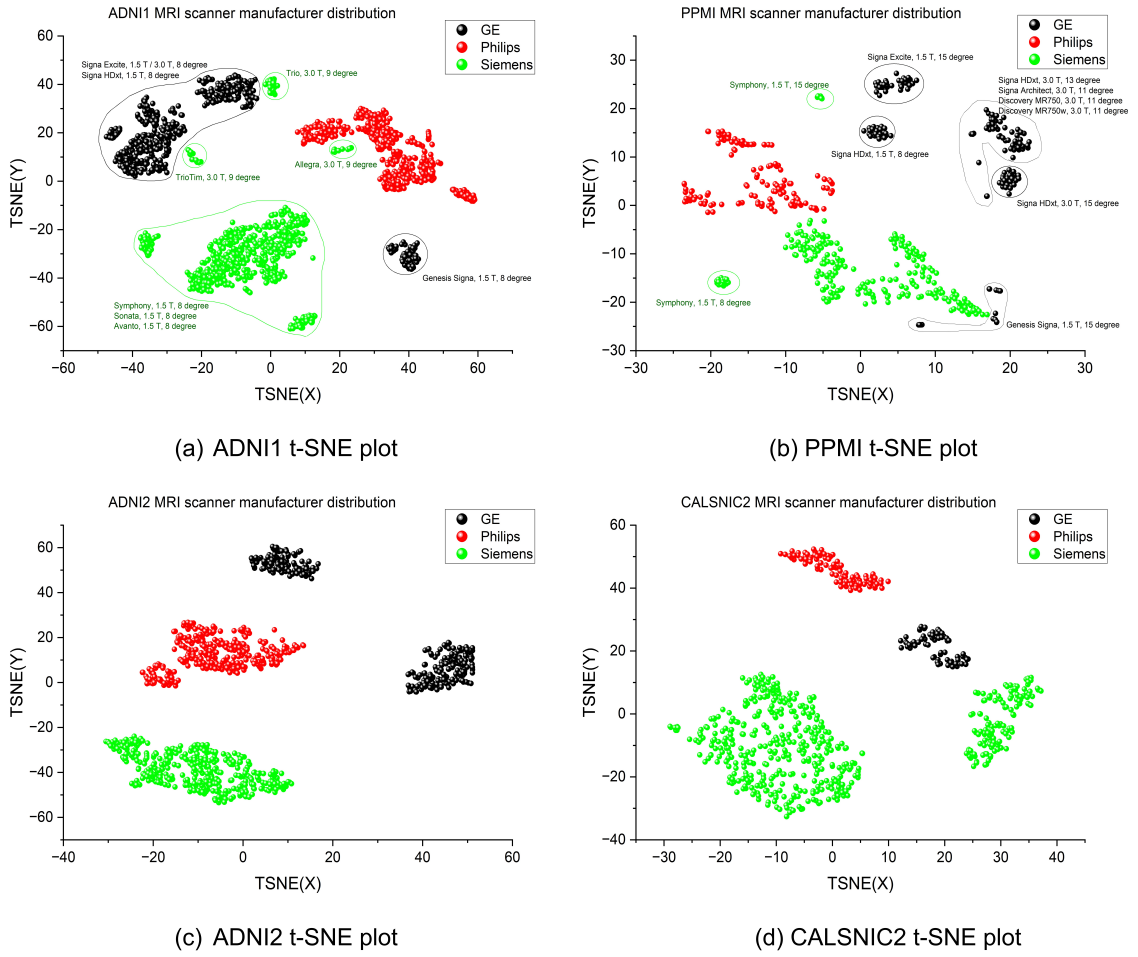


Figure 5.4: t-SNE plots for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets using the features generated by MRQy evaluation metrics. Different clusters are primarily formed based on the scanner manufacturer. In panels (a) and (b), bounding boxes are delineated, incorporating information about the scanner model, field strength, and flip angle. These annotations visually highlight their role in inducing domain shift within a dataset.

linear, graph-based dimension reduction method that project the high-dimensional feature space into a lower-dimensional space while preserving the distribution characteristics. The visualization of the t-SNE plots reveals that the proximity of grouped data primarily corresponds to the scanner manufacturer. Additionally, I observe further clustering within the same vendor, which can be attributed to variations in scanner models from the same manufacturer. Minor contributions to data clustering arise from variations in magnetic field strength and flip angles, as depicted by dif-

ferent bounding boxes in Fig. 5.4. Some 3D views of the t-SNE and UMAP plots are available on my GitHub project page at (https://github.com/rkushol/Effects-of-MRI-scanner-manufacturer).

5.2.5 Gender Classification

The task of gender classification (Male vs. Female) from MRI data is comparatively less intricate than the challenge of classifying different neurodegenerative diseases. In this context, I evaluate gender classification across the four previously mentioned datasets to assess performance variations among different scanner manufacturers. The outcomes of gender classification, achieved through distinct 3D CNN-based deep models (ResNet [135], MobileNetV2 [137], ShuffleNetV2 [141]), are presented in Table 5.2. For the ADNI1, ADNI2, and CALSNIC2 datasets, the aforementioned CNN methods achieve an average accuracy and F1-score of over 90%. Notably, in the PPMI dataset, using data from Siemens and GE also yields an average accuracy of around 90%, while using Philips data results in an approximate classification accuracy of 85%. Overall, there is no significant difference in performance among the scanner manufacturers in this classification task.

Table 5.2: Gender classification results for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets.

Scanner	**DL**	**ADNI1**		**ADNI2**		**PPMI**		**CALSNIC2**	
	models	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**	**F1-score**	**Acc**	**F1-score**
GE	ResNet	0.92	0.93	0.93	0.94	0.93	0.92	0.92	0.92
	ShuffleNetV2	0.95	0.94	0.92	0.93	0.89	0.90	0.96	0.96
	MobileNetV2	0.92	0.93	0.91	0.90	0.88	0.89	0.92	0.92
Siemens	ResNet	0.94	0.94	0.92	0.91	0.90	0.90	0.91	0.90
	ShuffleNetV2	0.94	0.93	0.97	0.95	0.94	0.92	0.92	0.92
	MobileNetV2	0.90	0.89	0.91	0.90	0.88	0.88	0.90	0.91
Philips	ResNet	0.92	0.91	0.90	0.90	0.86	0.87	0.95	0.94
	ShuffleNetV2	0.93	0.93	0.90	0.88	0.85	0.84	0.94	0.94
	MobileNetV2	0.90	0.89	0.92	0.93	0.84	0.83	0.93	0.92

5.2.6 Disease Classification

Classifying patients with neurodegenerative diseases such as AD, PD, or ALS from healthy controls using limited MRI data poses significant challenges due to the subtle structural changes present in the images. To enhance the reliability of my findings while maintaining balanced sample sizes across different scanner manufacturers, I leverage longitudinal data. However, a notable exception arises in the CALSNIC2 dataset, where the volume of data from GE and Philips scanners is comparatively smaller compared to that of the Siemens vendor. Moreover, I ensure that the data-splitting strategy avoids data leakage issues. This involves meticulously dividing the data based on individual subjects, preventing mixing the same participant's images in both training and testing processes, as illustrated in Fig. 5.5.

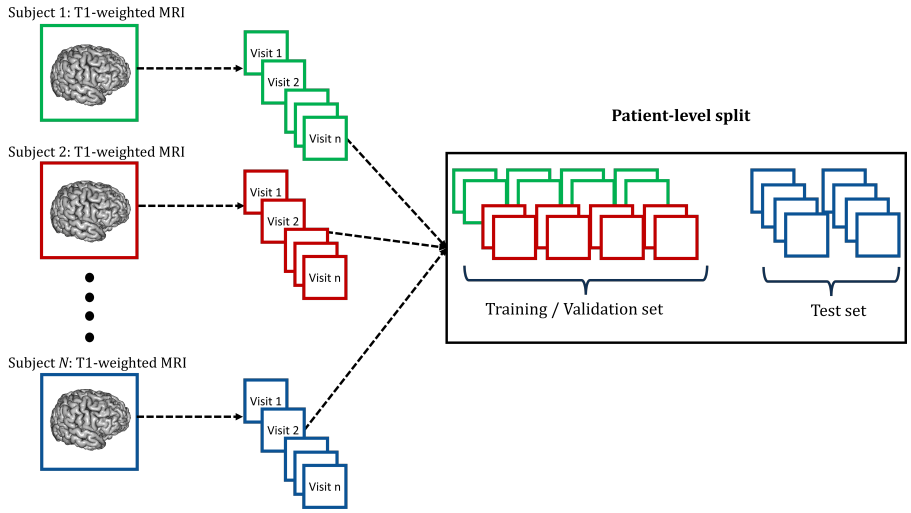


Figure 5.5: Patient-level split process for longitudinal data to train different DL models.

In the context of 2D frameworks, this practice is extended to ensure the integrity of slices within subjects across the test and training sets. Indeed, a recent study [147] discovered that many prior disease classification approaches did not follow a proper distribution of slices or subjects in their training or testing data. As a result, their reported outcomes present inaccurate and excessively optimistic classification accuracies. My analysis reveals that the ResNet (3D) and FCN (3D) models outperform

other 3D frameworks across various disease classification tasks. Similarly, in the case of 2D networks, the ResNet (2D) and ADDFormer (2D) models achieve better results compared to other 2D DL methods. Table 5.3 summarizes the classification results from these top-performing models.

Table 5.3: Different disease classification results based on scanner manufacturer with the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets. [G: GE, S: Siemens, P: Philips]

Scanner	**DL models**	**AD vs. CN**				**MCI vs. CN**		**PD vs. CN**		**ALS vs. CN**	
		ADNI1		**ADNI2**		**ADNI1**		**PPMI**		**CALSNIC2**	
		Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
GE	ResNet (3D)	0.76	0.76	0.79	0.79	0.70	0.68	0.80	0.80	0.70	0.70
	FCN	0.84	0.83	0.84	0.83	0.74	0.74	0.84	0.83	0.75	0.74
	ResNet (2D)	0.81	0.79	0.81	0.81	0.71	0.70	0.79	0.79	0.71	0.70
	ADDFormer	0.86	0.85	0.89	0.88	0.75	0.73	0.88	0.87	0.82	0.79
Siemens	ResNet (3D)	0.77	0.78	0.80	0.82	0.71	0.71	0.66	0.66	0.71	0.72
	FCN	0.84	0.84	0.82	0.82	0.73	0.73	0.70	0.71	0.75	0.76
	ResNet (2D)	0.78	0.76	0.76	0.74	0.71	0.70	0.66	0.66	0.71	0.69
	ADDFormer	0.88	0.88	0.86	0.85	0.71	0.72	0.72	0.71	0.78	0.79
Philips	ResNet (3D)	0.75	0.73	0.83	0.82	0.66	0.66	0.77	0.76	0.70	0.69
	FCN	0.84	0.83	0.86	0.85	0.71	0.70	0.80	0.80	0.73	0.72
	ResNet (2D)	0.74	0.73	0.79	0.78	0.67	0.66	0.74	0.73	0.70	0.67
	ADDFormer	0.85	0.85	0.91	0.90	0.71	0.71	0.82	0.82	0.79	0.79
All samples (G+S+P)	ResNet (3D)	0.76	0.78	0.79	0.80	0.71	0.69	0.76	0.77	0.72	0.72
	FCN	0.84	0.85	0.85	0.84	0.77	0.75	0.78	0.78	0.74	0.75
	ResNet (2D)	0.77	0.76	0.78	0.78	0.72	0.71	0.73	0.72	0.72	0.70
	ADDFormer	0.88	0.88	0.89	0.89	0.76	0.75	0.80	0.79	0.81	0.81
One-third samples (G+S+P)	ResNet (3D)	0.72	0.70	0.78	0.77	0.66	0.68	0.73	0.74	0.67	0.67
	FCN	0.80	0.80	0.80	0.81	0.70	0.68	0.74	0.74	0.71	0.70
	ResNet (2D)	0.74	0.72	0.75	0.75	0.66	0.65	0.70	0.70	0.67	0.68
	ADDFormer	0.79	0.80	0.80	0.79	0.68	0.68	0.75	0.74	0.74	0.75

ADNI1

Firstly, the independent evaluation of AD classification performance across the three manufacturers yields very close accuracy results. The classification accuracy of the top-performing model falls within the range of 85%-88%. Secondly, comparable accuracy is achieved when combining data from all manufacturers, resulting in a sample size approximately three times larger than that of each individual vendor. How-

ever, when equalizing the total sample size to that of a single manufacturer (approximately one-third of the total samples), a noticeable decline in performance is observed. Thirdly, among the 3D frameworks, the customized FCN model achieves the highest score, while the ADDFormer model outperforms all others in terms of classification accuracy. On the other hand, a similar conclusion is depicted for the intermediate stage of AD, known as the MCI vs. CN classification task, except that the overall accuracy decreases from all angles.

ADNI2

The classification accuracy of ADNI2 slightly surpasses that of ADNI1. Among the three manufacturers, utilizing data from Philips scanners yields slightly better performance compared to data from GE or Siemens. The range of the best model's classification accuracy falls between 86% and 91%. Upon merging data from all manufacturers, which increases the sample size to approximately three times that of individual vendors, the achieved accuracy remains consistent. However, performance experiences a noticeable decline when the sample size is reduced to that of a single manufacturer, accounting for roughly one-third of the total samples. Once again, among the 3D frameworks of DL models, both the ResNet and the custom-made FCN model achieve better results. In contrast, within the group of 2D methods, the ADDFormer model stands out for achieving the highest classification accuracy.

PPMI

In the PD vs. CN classification task, a few control samples from the ADNI2 dataset are added to ensure a balanced sample size of patients and healthy controls across all three manufacturer groups, thus mitigating the class imbalance issues. Notably, the FCN and ADDFormer custom-made models also demonstrate strong performance when compared to other fundamental CNN-based methods. The ShuffleNet achieves better outcomes in certain cases within the group of 3D frameworks. The range

of the best model's classification accuracy spans from 72% to 88%. Comparable classification results are observed whether the data originates from GE or Philips scanners. However, the outcomes using data from Siemens are comparatively poor. This discrepancy could be due to sharing a small number of healthy control samples from the ADNI2 dataset, whereas the GE or Philips group shares a large number of control samples from the ADNI2. Likewise, employing a total sample size equivalent to that of an individual manufacturer (approximately one-third of the total samples) leads to a noticeable decline in performance.

CALSNIC2

The classification task involving ALS patients vs. healthy controls within the CALSNIC2 database presents an even greater challenge compared to AD classification. All three manufacturers exhibit similar average classification accuracy. However, the performance of data originating from Siemens scanners is notably more reliable due to the inclusion of large samples from multiple centers. The range of the best model's classification accuracy falls between 78% and 82%. The accuracy remains consistent when the data from all manufacturers are combined. Conversely, the performance experiences a noticeable decline when the sample size from the Siemens manufacturer is reduced to one-third. The number of scans from GE and Philips scanners remains unchanged, as their original sizes are already limited. Among the DL models in both 3D and 2D frameworks, the ADDFormer model once again stands out for its highest classification accuracy.

Cross-validation

This section examines the consequences of introducing a change in the test set data by employing a different manufacturer. The left panel of Table 5.4 illustrates the classification results for this cross-domain validation using the four top-performing DL models described earlier. In this experimental setup, data originating from a

specific manufacturer is utilized as the training domain, while the remaining two serve as the test domains. When comparing these findings with the results presented in Table 5.3, it becomes evident that a significant drop in accuracy is observed across all datasets in Table 5.4. These outcomes further confirm the presence of a substantial domain shift inherent within the MRI data acquired from different manufacturers.

Table 5.4: The cross-domain intra-study disease classification accuracy before and after voxel-wise ComBat harmonization for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets. [G: GE, S: Siemens, P: Philips]

Dataset	**Training data**	**Testing data**	**Classification Acc with different DL models**							
			Results before harmonization				**Results after harmonization**			
			ResNet (3D)	FCN (3D)	ResNet (2D)	ADDFo-rmer(2D)	ResNet (3D)	FCN (3D)	ResNet (2D)	ADDFo-rmer(2D)
ADNI1	GE	P + S	0.75	0.80	0.75	0.86	0.75	0.78	0.74	0.76
AD vs.	Philips	G + S	0.69	0.74	0.68	0.71	**0.70**	0.72	**0.69**	**0.73**
CN	Siemens	G + P	0.72	0.77	0.74	0.79	0.71	0.71	0.70	0.68
ADNI2	GE	P + S	0.71	0.76	0.72	0.80	0.69	0.71	0.69	0.68
AD vs.	Philips	G + S	0.71	0.74	0.71	0.75	0.63	0.65	0.66	0.63
CN	Siemens	G + P	0.75	0.77	0.77	0.83	0.68	0.70	0.67	0.69
ADNI1	GE	P + S	0.66	0.71	0.66	0.71	0.64	0.67	**0.68**	0.70
MCI vs.	Philips	G + S	0.60	0.67	0.62	0.64	**0.61**	0.62	0.59	0.64
CN	Siemens	G + P	0.65	0.67	0.64	0.66	0.65	0.64	0.63	0.65
PPMI	GE	P + S	0.62	0.63	0.60	0.63	0.60	0.62	0.59	0.56
PD vs.	Philips	G + S	0.65	0.66	0.63	0.67	0.60	0.65	0.59	0.59
CN	Siemens	G + P	0.56	0.61	0.60	0.60	**0.59**	**0.63**	**0.62**	**0.67**
CALSNIC2	GE	P + S	0.57	0.56	0.56	0.61	0.57	**0.57**	0.55	0.55
ALS vs.	Philips	G + S	0.59	0.59	0.60	0.62	0.56	0.58	**0.61**	0.62
CN	Siemens	G + P	0.61	0.63	0.65	0.68	0.59	**0.65**	0.65	**0.71**

5.2.7 ComBat Harmonization Effects

Initially, I evaluate the outcomes of a modified ComBat-based method known as ComBat-generalized additive model (ComBat-GAM), specifically designed to address site effects in multi-site neuroimaging datasets [102]. ComBat-GAM is the only publicly available package that directly handles 3D NIFTI images as input, accessible at (https://github.com/rpomponio/neuroHarmonize). This technique successfully estimated age-related volume differences within a large-scale multi-center dataset, seg-

menting each MR image into 145 ROIs. However, my analysis does not yield promising outcomes when harmonizing entire 3D MRI data, as opposed to limited features extracted from MR images. Appendix Fig. A.3 provides an example of a 2D axial brain slice before and after harmonization using the ComBat-GAM method from the CALSNIC2 dataset. The output image exhibits undesirable artifacts and blurriness, with distinct brain tissue sections showing abnormal patterns of intensity shift compared to the input image. This disrupts the structural integrity of gray and white matter. As a result, I abstain from performing classification tasks using these undesirable resultant images generated by the ComBat-GAM approach. Subsequently, I apply the standard ComBat method to the multi-center datasets, utilizing the official implementation available at (https://github.com/Jfortin1/ComBatHarmonization). A minor adjustment is made to the original implementation to enable voxel-level harmonization instead of feature-level harmonization, treating each scanner manufacturer as an individual site. From a visual perspective, the outcomes produced by the standard ComBat method closely resemble the original images, with minor changes evident in cortical regions, as depicted in Fig. 5.6. Thus, I harmonize the datasets using the standard ComBat and utilize the harmonized images for the cross-domain classification context. The classification results following the ComBat harmonization are presented in the right panel of Table 5.4. Unfortunately, the harmonized images generated by the standard ComBat method demonstrate weakness in enhancing the classification accuracy in most cases (exceptions are shown in bold in Table 5.4). The potential reason behind these failures could be that ComBat-based harmonization techniques are inappropriate for image/voxel-level harmonization. Successful ComBat-based applications reported in prior studies have predominantly focused on limited feature-level harmonization. Notably, during the execution of both ComBat-based strategies, I incorporate age and sex as covariates to ensure the preservation of this biological information throughout the harmonization process.

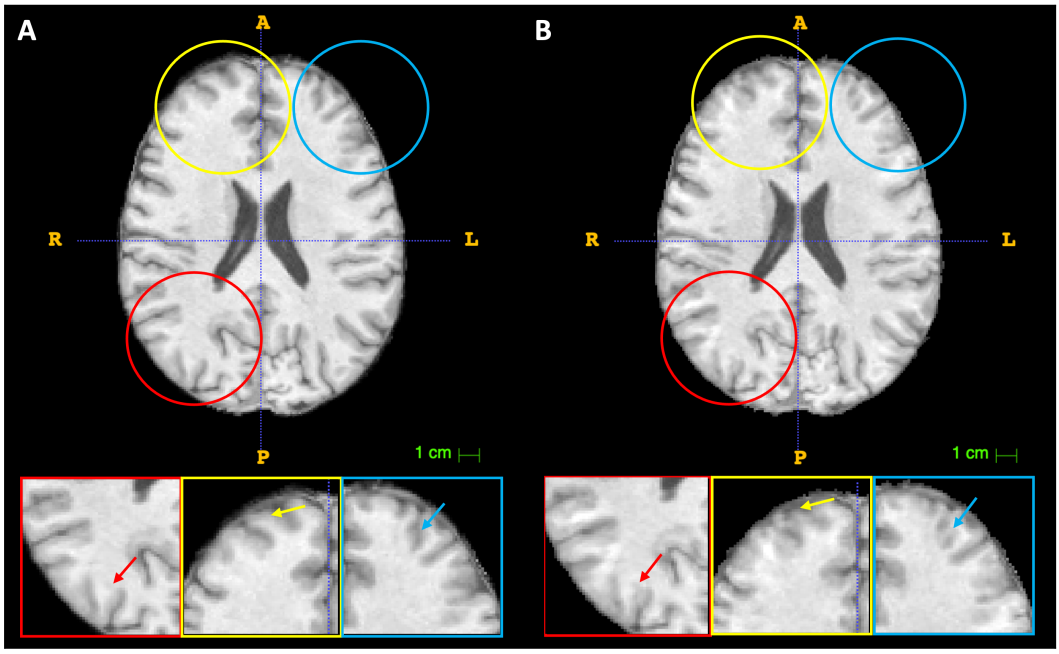


Figure 5.6: Minor changes in voxel-wise ComBat harmonization using structural MRI. A) One 2D axial slice of preprocessed 3D T1-weighted MR image of CALSNIC2 dataset before harmonization, B) corresponding slice after harmonization. The red, yellow, and blue arrows point to the regions with manipulated structures, including the disappearance of minor details resulting from the ComBat harmonization.

5.2.8 Quality Evaluation of Scanners Data

Alongside manual inspection, I utilize the quality control tool MRQy [57] to verify the quality of each MR image. The MRQy tool offers a comprehensive array of quality-related metrics, including peak signal-to-noise ratio (PSNR), contrast-to-noise ratio (CNR), coefficient of variation of the foreground patch (CVP) to address shading artifacts, coefficient of joint variation (CJV) to quantify aliasing and inhomogeneity artifacts between foreground and background, and entropy focus criterion (EFC) to detect motion artifacts. The user-friendly interface of MRQy greatly simplifies the process of identifying outliers or inconsistencies within a dataset. Table 5.5 presents an illustrative comparison of the diverse quality metrics obtained by averaging all samples for each scanner manufacturer.

Table 5.5: The quality evaluation of MRI data with MRQy for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets.

Dataset	Quality metrics	MRI scanner manufacturer		
		GE (Mean±Std)	Siemens (Mean±Std)	Philips (Mean±Std)
ADNI1	PSNR ↑	15.69±2.8	16.89±1.1	**18.23±1.6**
	CNR ↑	21.18±9.0	19.41±5.0	**50.16±18.9**
	CVP ↓	**0.36±0.1**	0.42±0.1	0.41±0.1
	CJV ↓	0.88±0.2	**0.85±0.1**	1.22±0.3
	EFC ↓	**2.49±0.4**	2.67±0.2	2.60±0.3
ADNI2	PSNR ↑	16.96±1.0	15.37±0.9	**17.16±1.1**
	CNR ↑	17.64±17.3	**34.09±6.5**	12.31±2.0
	CVP ↓	0.46±0.1	**0.41±0.1**	0.51±0.1
	CJV ↓	1.59±2.6	**0.91±0.1**	1.48±0.5
	EFC ↓	**1.89±0.1**	2.94±0.1	2.17±0.2
PPMI	PSNR ↑	13.65±1.5	14.42±1.2	**17.31±3.6**
	CNR ↑	29.29±25.3	**34.99±12.99**	16.33±5.7
	CVP ↓	**0.39±0.1**	0.41±0.1	0.47±0.1
	CJV ↓	**0.84±0.2**	**0.84±0.1**	0.95±0.3
	EFC ↓	24.02±13.1	4.04±1.8	**3.36±1.5**
CALSNIC2	PSNR ↑	**14.98±0.9**	12.59±1.6	11.82±1.0
	CNR ↑	16.45±4.3	**70.05±24.8**	10.54±2.4
	CVP ↓	0.41±0.1	**0.37±0.1**	0.46±0.1
	CJV ↓	**0.72±0.1**	0.84±0.1	0.82±0.1
	EFC ↓	10.1±3.9	8.19±3.0	**2.61±0.2**

5.3 Discussion

The reproducibility of MRI research continues to be challenging, particularly when data is influenced by scanner effects, a type of non-biological variation originating from various image acquisition protocols. After demonstrating significant distinguishable imaging characteristics present in data derived from distinct scanner manufacturers, I explore its consequences for different disease classification tasks using several prominent 2D and 3D DL models. The specialized FCN model consistently outperformed other 3D classification frameworks in most disease classification scenarios. A notable advantage of 3D frameworks lies in their ability to process the entire brain as input, eliminating the need for prior knowledge in selecting specific slices for feature extraction. However, 3D DL methods tend to lack the utilization of pre-trained networks

through transfer learning. In contrast, 2D frameworks necessitate the careful selection of relevant 2D slices based on prior knowledge. Additionally, the 2D DL models leverage the transfer learning property by utilizing pre-trained models with a massive 2D imaging dataset like ImageNet. Overall, the ADDFormer network demonstrates the best performance in this study, leveraging the power of the ViT architecture by integrating spatial and frequency domain features in a novel manner.

The preprocessing steps applied to our original T1-weighted MR images involve state-of-the-art algorithms and can be easily replicated using open-source tools. After experimenting with a straightforward classification task of differentiating sex (male vs. female) using the original MRI data, I move on to more sophisticated neurodegenerative disease classification tasks. Based on the results obtained from different DL models, the most challenging classification task is distinguishing between MCI and CN groups. This finding aligns with prior studies, which have also reported lower accuracy in this specific classification [83]. Notably, some investigations have further subdivided MCI into progressive (pMCI) and stable (sMCI) subgroups, achieving improved results through such stratification [148].

The next challenging task is the classification of PD vs. CN. One critical factor that makes this classification task difficult is the heterogeneous nature of the dataset. The PPMI dataset encompasses 21 different centers [129], a characteristic evident in Fig. 5.4 (b). As a result, a decline in performance is anticipated in DL models if the test set contains data from a particular center, while the corresponding center's data is either insufficient or entirely missing in the training set. For the same reason, tasks such as scanner vendor and gender classification might yield lower accuracy with the PPMI dataset compared to others. Lastly, the classification task of distinguishing between ALS patients and healthy controls also presents challenges due to the insignificant structural changes in MRI data compared to the control group.

Chapter 6

Spatial and Frequency Fusion Transformer

This chapter extends the idea of ADDFormer architecture more comprehensively for the challenging task of amyotrophic lateral sclerosis (ALS) classification. This study introduces an effective and robust transformer-based framework titled $SF^2Former$ (Spatial and Frequency Fusion transFormer) for the classification of ALS subjects and healthy controls using multi-modal brain MRI data (i.e., T1-weighted, R2*, FLAIR). The source code is available at https://github.com/rkushol/ADDFormer.

6.1 Proposed Method

6.1.1 Overview

The proposed framework encompasses simple preprocessing steps of the raw MRI data using FreeSurfer [82] and FSL [132], selection of a fixed range of 2D coronal slices, combination of features from two transformer networks in the spatial and Fourier domains, and majority voting on the predictions of individual slices to determine the final classification result. Figure 6.1 depicts the overall workflow of the proposed phases.

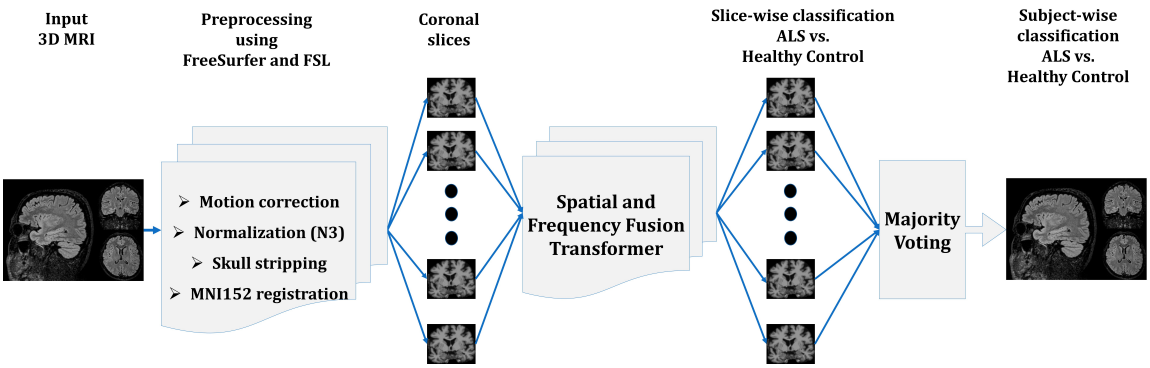


Figure 6.1: The overall workflow of the proposed stages.

6.1.2 Preprocessing

A straightforward, fast, and easy-to-perform preprocessing pipeline has been followed on the original 3D brain MRI data to make it suitable for deep models. Several common preprocessing operations are conducted using the FreeSurfer program, including motion correction, skull stripping, and non-parametric non-uniform intensity normalization (N3) (`Command: recon-all -subject subjectname -i input-file.nii -autorecon1`). Subsequently, registration to the MNI-152 standard space is performed using the FSL `flirt` function. After the reconstruction of the original images, the resultant image dimension is a matrix of size $182 \times 218 \times 182$ with a voxel size of $1 \times 1 \times 1\ mm^3$. I utilize an eight-core CPU platform that leverages parallel processing, resulting in an average processing time of approximately 5 minutes per scan for the preprocessing steps. Without parallelization, processing each subject individually takes approximately 15 minutes per scan. This efficient preprocessing pipeline ensures that the MRI data is prepared for subsequent analysis with deep models.

6.1.3 Slice Selection

After conducting empirical analysis, the optimal performance is found by manipulating the coronal slices among the three planes (coronal, sagittal, and axial) of the 3D MRI scans. To identify a potential region of meaningful slices, I extensively explore various combinations of slices through training and testing the network. The detailed

outcomes of the experiments with various slice clusters are given in section 6.2.7. Ultimately, 15 consecutive 2D images from the central section of the coronal plane are used to train the proposed framework. Expert observations suggest that this range of coronal slices effectively captures the CST, a prominent region of interest in ALS. It is essential to mention that the slices generated from the same subject are never used simultaneously in both the train or test sets. In other words, the proposed method follows a subject-level split protocol to avoid data leakage, which is further illustrated in Fig 6.2. Data leakage in a machine learning model refers to the inadvertent sharing of information between the test and training datasets, leading the model to already possess knowledge about certain aspects of the test data after training. Indeed, a recent study [147] revealed that many previous approaches to neurodegenerative disease classification did not adhere to proper slice division in their training or testing data, resulting in incorrect and excessively optimistic classification accuracies.

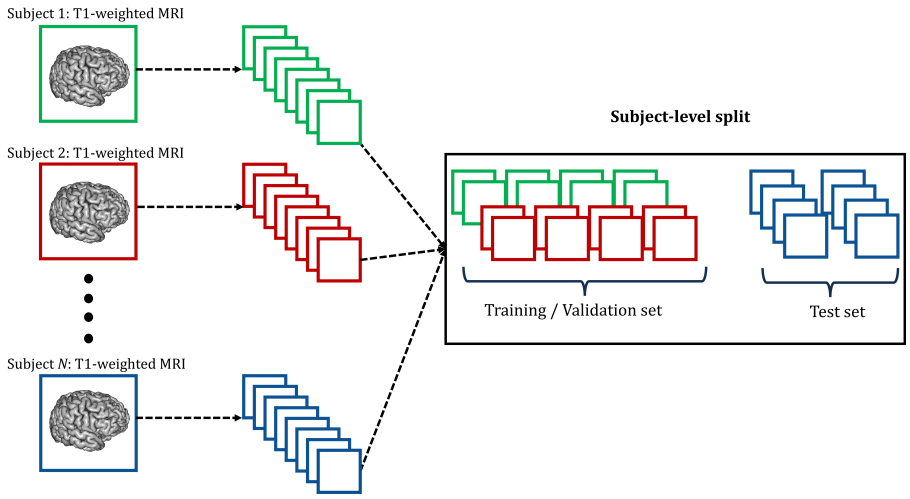


Figure 6.2: Subject-level split process of the data used to train the proposed model.

6.1.4 Architecture

Figure 6.3 depicts the overall architecture of the proposed $SF^2Former$ method, which integrates features from two vision transformer-based networks. One network is responsible for generating features from the spatial domain, and the other is capable of

developing features from the frequency domain.

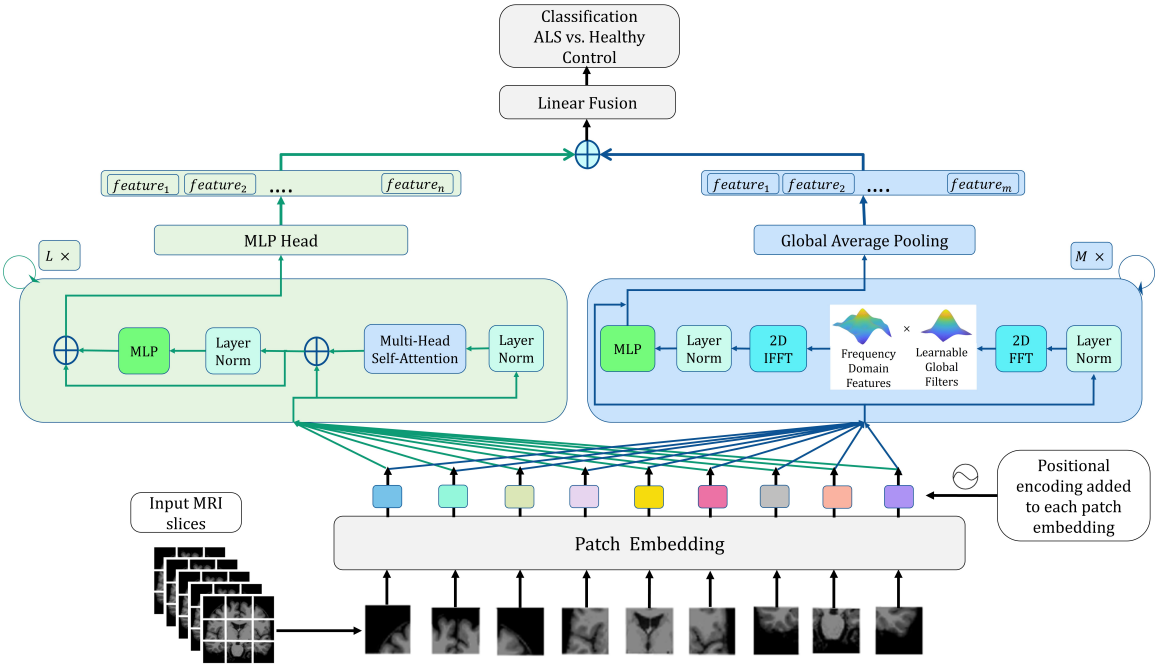


Figure 6.3: The overall architecture of the proposed $SF^2Former$ framework. The left branch of the methodology encodes features from the spatial domain, whereas the right segment encodes features from the frequency domain. Finally, the linear fusion module incorporates the features to assemble the classification decision for each 2D slice.

The ViT is the first successful method to adopt the transformer architecture and has achieved state-of-the-art performance in many computer vision tasks. Unlike other deep learning-based approaches that extract features from images using convolution, which face challenges in capturing long-term dependencies, ViT employs self-attention (SA) to overcome this limitation. However, SA involves computationally expensive matrix multiplications. To mitigate this, ViT embeds the image into patch tokens and uses these tokens as input to reduce the computational complexity. The left branch of the proposed architecture, similar to ViT, consists of alternating layers of multiheaded self-attention (MSA) and a Multilayer perceptron (MLP) block with two layers of Gaussian Error Linear Unit (GELU) [149] non-linearity on top of the encoder. Layer normalization (LN) is applied before each block, and residual connections are adjusted after each block. L represents the number of transformer

encoder layers which is 12 in our case.

To begin with, I have 2D slices with a spatial resolution of (H, W) and C channels. As an input to the transformer, I reshape the image $x \in \mathbb{R}^{H\times W\times C}$ into a series of 2D patches $x_p \in \mathbb{R}^{N\times(P^2.C)}$. Here (P, P) represents the size of each patch, and I have $N = HW/P^2$ number of total patches, which serves as the input sequence length for the transformer. Now, I flatten the patches and project them to D dimensions using a trainable linear projection. This is necessary because the transformer maintains a constant latent vector size D throughout its layers. The output of this mapping is known as patch embeddings, as shown in Fig 6.3. To preserve positional information, position embeddings $E_{pos} \in \mathbb{R}^{N\times D}$ are appended to the patch embeddings. I employ learnable 1D position embeddings, following the work [38]. The value of z at different layers and positions can be expressed as follows:

$$z_0 = [x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos} \qquad E \in \mathbb{R}^{(P^2.C)\times D} \tag{6.1}$$

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \qquad l = 1 \cdots L \tag{6.2}$$

$$z_l = MLP(LN(z_l')) + z_l' \qquad l = 1 \cdots L \tag{6.3}$$

The MSA module computes the relationship between pairs of tokens to produce the attention map A using an SA layer. Given an input sequence $z \in \mathbb{R}^{N\times D}$, the SA module linearly projects z into three embeddings: query q, key k, and value v, represented as $[q, k, v] = zU_{qkv}$, where $U_{qkv} \in \mathbb{R}^{D\times 3D_h}$. Subsequently, the attention map A is estimated by taking the dot product between the query q and key k, which is used to weight the value embedding v as follows:

$$SA(q, k, v) = Av = softmax(\frac{qk^\top}{\sqrt{D_h}})v \tag{6.4}$$

$$MSA(q, k, v) = Concat[SA_1(q, k, v), \cdots, SA_i(q, k, v)]U_{msa} \tag{6.5}$$

The MSA module's output involves performing i SA operations in parallel where weight matrix $U_{msa} \in \mathbb{R}^{i.D_h\times D}$. The dimension D_h corresponds to the size of each attention head, and the scaling factor $\sqrt{D_h}$ is used to stabilize the gradient during

training.

There is a close relationship between MRI data acquisition and the frequency domain. The mathematical formation of raw MRI data is initially performed in the frequency domain before being converted to the spatial domain for visual interpretation. This connection also motivates me to extract features in the frequency domain. However, directly involving a deep model in the frequency domain can be computationally expensive, especially when dealing with high-resolution images. The GFNet [87] utilizes a similar downsampling design as the ViT and replaces the SA layer with the Fast Fourier Transform (FFT) applied to the embedded patch tokens. The right-side branch of the proposed architecture introduced in Fig 6.3 essentially follows the concept of GFNet. The primary objective of this network is to learn the frequency domain interactions among different spatial positions. Another notable difference between GFNet and ViT is the use of global average pooling in the final feature map as an alternative to the extra class embedding head. GFNet accepts non-overlapping $H \times W$ patches as input and flattens them into $L = HW$ tokens with a dimension of D. Each spatial domain token $x \in \mathbb{R}^{H \times W \times D}$ transformed by 2D FFT generates a complex tensor X in the frequency domain as:

$$X = 2D\ FFT[x] \in \mathbb{C}^{H \times W \times D}. \tag{6.6}$$

Then, I modulate X (the spectrum of x) with respect to a learnable filter K in the form of element-wise multiplication and can be expressed as:

$$\tilde{X} = X \odot K. \tag{6.7}$$

The parameter K is a global filter representing an arbitrary frequency-domain filter with the same dimension as X. Ultimately, the modulated spectrum $\tilde{X}$ is transformed back to the spatial domain using the inverse FFT (iFFT), and the tokens are updated as:

$$x = 2D\ iFFT\ [\tilde{X}]. \tag{6.8}$$

The other components, such as layer norm and MLP, used in the diagram for GFNet, are identical to those in ViT. Among the different variants of GFNet, I adopt the transformer-style GFNet-B version, which consists of 19 layers/depth and has an embedding dimension of 512. Therefore, in my presented diagram of Fig. 6.3, the values of m and M become 512 and 19, respectively. To leverage information from both spatial and frequency domains, I propose a new linear fusion block to combine the features extracted from ViT and GFNet. A new linear head with a dimension equal to the joint embedding dimensions of ViT and GFNet is constructed, where the input comes from the final layers of these two networks as a form of concatenation. The output from the linear fusion block containing the merged features of the spatial and frequency domain is used to carry out the classification decision. The loss function we operate throughout our model is the cross-entropy (CE) loss function.

6.1.5 Majority Voting

The rationale behind applying majority voting in the proposed methodology is that it is unlikely for all the disease-affected tissues or areas to be present in all the chosen slices. In other words, automatically identifying 2D slices that consistently capture distinct clinical characteristics for every subject without manual intervention is not feasible. To address the possibility of false-positive responses from these insignificant slices, I employ the concept of majority voting at the last stage. The final classification of an individual sample is determined based on the class that has the highest number of occurrences within the given range of slices. The effectiveness of this majority voting scheme in enhancing classification accuracy is further reported in the ablation study 6.2.4 section.

6.2 Experiments

6.2.1 Dataset

Neuroimaging data are obtained from two independent datasets of the CALSNIC[1] [131] study. CALSNIC is a longitudinal, multi-center, and multi-modal study in which 3T MRI scans are acquired using scanners from three different manufacturers: GE Healthcare, Philips, and Siemens. The data used in the experimentation with CALSNIC1 is accumulated from five centers (i.e., Calgary, Edmonton, Toronto, Vancouver, and Montreal), whereas CALSNIC2 data comprises seven different centers (i.e., Calgary, Edmonton, Toronto, Quebec, Miami, Utah, and Montreal). To avoid potential data leakage issues, I only consider MRI data from participants with a certain visit (baseline) in our experiments. Due to a shortage of data in CALSNIC1, the FLAIR and R2* modalities are only considered from the CALSNIC2 dataset. A summary of the demographic information for both datasets is provided in Table 6.1.

6.2.2 Implementation

The proposed framework is implemented using PyTorch [146] and runs on a server with 4 NVIDIA 2080 Ti GPUs. The coding follows the publicly available implementation of the ViT[2] and the GFNet[3]. To ensure robustness during training, I apply data augmentation techniques such as random rotations and flipping to create diverse training batches. Each coronal slice is resized to a dimension of 224×224. After normalization, the pixel intensity values are scaled to the range of [0, 1]. I employ the SGD optimizer to train all the networks with a momentum value of 0.9. I choose an initial learning rate of 0.001 and decay the rate to 10^{-5} using the cosine schedule. A batch size of 16 and a total number of 150 epochs are used for both transformer networks. A concise summary of the network parameter details of our

[1](https://calsnic.org/)
[2](https://github.com/jeonsworld/ViT-pytorch)
[3](https://github.com/raoyongming/GFNet)

Table 6.1: Demographic details of T1-weighted MR images for the CALSNIC1 and CALSNIC2 datasets. ALSFRS-R = The Revised Amyotrophic Lateral Sclerosis Functional Rating Scale, S.D. = Standard Deviation.

Participant characteristics	**CALSNIC1**			**CALSNIC2**		
	ALS patients	**Healthy controls**	**p-value**	**ALS patients**	**Healthy controls**	**p-value**
Subjects	61	59	-	116	116	-
Sex: Male/Female	36/25	27/32	0.15	76/40	57/59	0.01*
Age (years)						
Mean ± S.D.	58.4 ± 10.7	54.0 ± 10.2	0.02*	60.1 ± 10.1	56.3 ± 10.6	0.005*
Median	57.0	55.0	-	60.9	58.7	-
Range	33.0 - 86.0	25.0 - 69.0	-	25.6 - 83.4	25.8 - 77.0	-
ALSFRS-R score						
Mean ± S.D.	39.2 ± 5.0	-	-	37.3 ± 7.0	-	-
Median	40.0	-	-	39.0	-	-
Range	22.0 - 47.0	-	-	7.0 - 47.0	-	-
Symptom duration (months)						
Mean ± S.D.	16.1 ± 10.5	-	-	22.6 ± 13.2	-	-
Median	13.3	-	-	19.3	-	-
Range	4.0 - 54.8	-	-	2.6 - 59.8	-	-

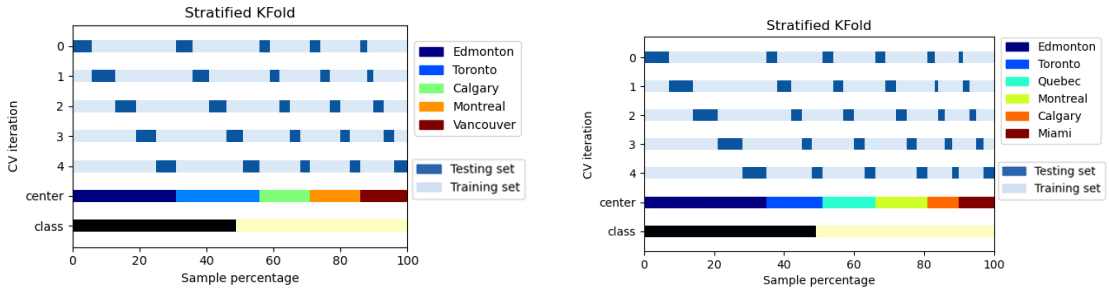
proposed framework is given in Table 6.2. The final accuracy and the values of hyper-parameters reported in the study are attained from five-fold cross-validation (CV). Specifically, we design a variation of the Stratified KFold approach where each fold is conditioned on a similar share of data from all the available centers. This ensures that each fold will incorporate approximately the same percentage of samples from their respective centers. For a better understanding of the data distribution for model training in CALSNIC1 and CALSNIC2, Fig. 6.4 is presented below. The data split ratio of train: validation: test data is 7:1:2 in each fold. The training time of the proposed approach is approximately 6 hours when operating on a single GPU with 12GB memory.

6.2.3 Evaluation Metrics

Commonly used statistical metrics, such as accuracy, sensitivity, specificity, precision, and F1-score are used to assess the classification performance of the pro-

Table 6.2: The summary of parameter details of the proposed framework.

Parameters	**Left branch/ spatial domain**	**Right branch/ frequency domain**
Image dimension	224×224	224×224
Patch size	16×16	16×16
Number of layers/ depth	$L = 12$	$M = 19$
Embedding dimension	$n = 768$	$m = 512$
Activation function	GELU	GELU
MLP dimension	3072	2048
Dropout rate	0.1	0.25



(a) CALSNIC1 (b) CALSNIC2

Figure 6.4: Stratified five-fold cross-validation (CV) designed for CALSNIC datasets. The row labelled `'class'` indicates the percentage of ALS patients and healthy controls, the number of which is similar and balanced in both datasets. Next, the row tagged `'center'` shows the percentage of participants in the corresponding dataset from available centers. The five rows above `'center'` show training and test data distribution with five iterations of CV. Each iteration involves a similar proportion of samples from each center.

posed method. These metrics are defined in terms of four values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Sensitivity (SEN), also known as recall, is the capability of a test to correctly identify patients with a disease and is defined as $SEN = \frac{TP}{TP+FN}$. Specificity (SPE), also known as the true negative rate, is the ability to determine individuals without the disorder and is expressed as $SPE = \frac{TN}{TN+FP}$. The positive predictive value or precision (PRE) reflects the proportion of relevant items, with high precision indicating that the algorithm generates significantly more relevant outcomes than irrelevant ones ($PRE = \frac{TP}{TP+FP}$). Accuracy (ACC) represents the fraction of correctly identified

subjects to the total number of samples in a specific database and is computed as $ACC = \frac{TP+TN}{TP+TN+FP+FN}$. The F1-score is the harmonic mean of precision and recall, defined as $F1 - score = 2 \times \frac{precision \times recall}{precision + recall}$.

6.2.4 Ablation Study

To demonstrate the effectiveness of different components and techniques used in the proposed method, an ablation study is summarized in Table 6.3. The values of different hyperparameters and the choice of slice selection remain the same for this experiment. I remove one particular function from the proposed framework to evaluate the impact of that function in the first five settings. First, I examine the model without normalizing the pixel intensities of MRI scans. Normalization of the images noticeably boosts the overall accuracy. Next, I assess the importance of data augmentation and observe a considerable reduction in accuracy when data augmentation is not applied. Importantly, slice selection plays a vital role in the proposed technique. In particular, including most of the slices of a 3D brain MRI lowers the network's performance. In the experiment without slice selection, I chose 120 coronal slices out of 218 (from slice 51 to 170) as slices outside this range do not contain the brain and thus lack meaningful information. Another influential function of the proposed method is leveraging the transfer learning property from the enormous computer vision dataset ImageNet [37]. Instead of using the pre-trained weights for the ViT and the GFNet networks, training from scratch decreases the accuracy slightly.

Moreover, the idea of applying majority voting at the end of the framework helps to enhance the performance significantly. In other words, bypassing this operation reduces the performance drastically. Finally, I evaluate the performance of the proposed framework's two major building blocks: the ViT and the GFNet. When I apply them independently, they can correctly determine samples which are not identical. However, I achieve much-improved classification accuracy in the proposed architecture by fusing the strength of these two networks. While some other methods may

Table 6.3: Ablation study for ALS patients vs healthy controls classification on a particular fold for T1-weighted MR images of CALSNIC1 dataset.

Method	ACC	SEN	SPE	PRE	F1score
w/o normalization	0.800	0.750	0.889	0.923	0.828
w/o augmentation	0.760	**1.000**	0.667	0.59	0.700
w/o slice selection	0.680	0.647	0.750	0.846	0.734
w/o transfer learning	0.840	0.846	0.834	0.846	0.846
w/o majority voting	0.720	0.759	0.686	0.676	0.715
ViT only [38]	0.800	0.834	0.769	0.769	0.800
GFNet only [87]	0.840	0.800	0.917	0.923	0.857
Proposed method	**0.880**	0.813	**1.000**	**1.000**	**0.900**

outperform the proposed technique in terms of SEN, they often sacrifice scores in SPE and PRE. The proposed methodology maintains a balance among all metrics, which is also reflected in the F1-score.

6.2.5 Effects of Multi-center Study

This section highlights the effect of multi-center data or data acquired with multiple scanners on classification performance. Deep learning-based models face increased challenges when MRI data originates from different scanners [150]. Here, I present the classification results from three different setups using CALSNIC2 T1-weighted images. I select samples from two major recruiting centers, namely Toronto and Edmonton, where Siemens 3T Prisma model scanners are used. In the final set, I randomly collect samples from six centers equal to the largest center's dimension. In the first setup, I estimate the classification accuracy using data from the Toronto center, consisting of 15 healthy controls and 20 ALS patients. In the second setup, I evaluate the classification accuracy of samples obtained from the Edmonton center, which includes 46 normal controls and 35 ALS subjects. Finally, I randomly assemble MR images from six centers (i.e., Calgary, Edmonton, Toronto, Quebec, Miami, and Montreal) to create a mixed dataset with a sample size similar to that of the Edmonton center. Table 6.4 illustrates the classification scores, showing that accuracy is higher when the data originates from a single center or the same type of scanner with an

identical image acquisition protocol. On the other hand, when the data comes from multiple centers or scanners, the performance declines. Another observation is the significance of having more training data for deep models. By including all the samples from the CALSNIC2 dataset, the classification accuracy improves from approximately 77% to 82%.

Table 6.4: Showing the effects of the multi-center study tested on CALSNIC2 T1-weighted MR images.

Center	**Samples**	**CALSNIC2**				
		ACC	**SEN**	**SPE**	**PRE**	**F1-score**
Toronto	35	0.813	0.900	0.708	0.800	0.845
Edmonton	81	0.824	0.770	0.571	0.727	0.824
Multi-center	81	0.765	1.000	0.714	0.600	0.765

6.2.6 Effects of Different MRI Modalities

The proposed framework investigates the applicability of multiple MRI sequences in classifying ALS from healthy controls. Firstly, I consider T1-weighted images from the CALSNIC1 and CALSNIC2 datasets as structural metrics (i.e., volume) obtained from T1-weighted scans commonly used for neurodegenerative disorder classification. Secondly, I evaluate the performance of the R2* modality from the CALSNIC2 dataset, which has rarely been explored for ALS classification. Finally, I calculate the classification accuracy using the FLAIR imaging modality on the CALSNIC2 dataset. Table 6.5 presents the classification performance of using different MRI modalities. Among the five evaluation metrics, this study reveals that the R2* modality achieves slightly better results in terms of accuracy, specificity, precision, and F1-score. The classification accuracy of T1-weighted images is very close to that of R2* and achieves the highest sensitivity. However, the FLAIR modality shows slightly lower classification accuracy compared to the other two modalities.

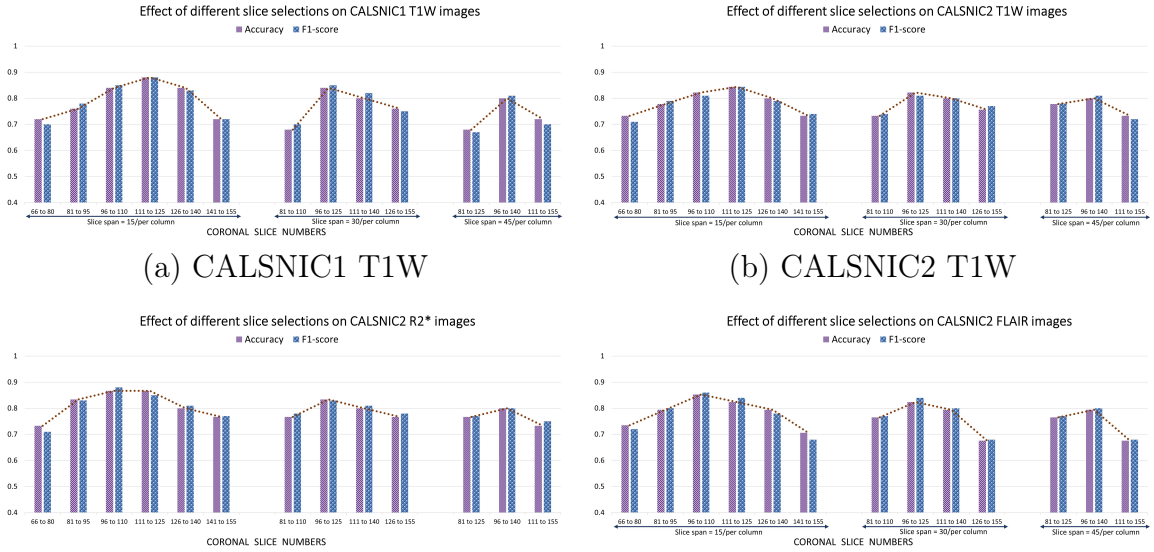
Table 6.5: Showing the classification results using different MRI modalities.

Modality (Dataset)	Samples	ACC	SEN	SPE	PRE	F1-score
T1-W (CALSNIC1)	120	0.816	**0.843**	0.812	0.800	0.815
T1-W (CALSNIC2)	223	0.818	0.824	0.815	0.800	0.811
R2* (CALSNIC2)	148	**0.820**	0.790	**0.875**	**0.880**	**0.829**
FLAIR (CALSNIC2)	168	0.806	0.807	0.824	0.812	0.803

6.2.7 Effects of Slice Selection

After applying the FreeSurfer *autorecon*1, FSL *flirt*, and resize commands, the reconstructed image size becomes $224 \times 218 \times 224$. As a result, each coronal slice (total 218) has a dimension of 224×224. From a 2D slice view perspective, the initial and final parts of the volume do not contain significant information. In other words, most of the tissues or important structural information can be found in the central part of the volume. Therefore, I explore the effectiveness of a wide range of slices to investigate which part of the volume provides the best performance, and the results are demonstrated in Fig 6.5. I start with an interval of 15 consecutive slices from the central slice location and analyze 45 slices in the forward and backward directions. Subsequently, I experiment with different combinations of successive slices within these 90-slice spans, such as 30 or 45 slices.

After careful observation, the best performance is found within the slice range of 111 to 125 for the T1-weighted images. The slice span of 96 to 110 demonstrates results closest to the optimal performance. Increasing the number of slices for training does not improve the performance, as noticed in each chart's middle and right segments in Fig 6.5. Moreover, increasing the number of slices for model training noticeably accelerates the overall training time. However, the best classification accuracy for the R2* maps and FLAIR images is achieved within the slice range of 96 to 110. The slices from 111 to 125 also perform comparably well in classifying ALS patients using R2* and FLAIR images. The anatomical features expected in different slice ranges can be perceived in Fig. 6.6.



(a) CALSNIC1 T1W

(b) CALSNIC2 T1W

(c) CALSNIC2 R2*

(d) CALSNIC2 FLAIR

Figure 6.5: Showing the classification accuracy effects on different ranges of coronal slice selections for each MRI modality used in the study.

6.2.8 Comparison

I compare the proposed network with popular deep learning-based 2D and 3D architectures and highlight the results in Table 6.6. Additionally, I reproduce the classification accuracy of a previous state-of-the-art work, M-CoHOG [13], for the CALSNIC1 dataset. Due to the manual slice selection process required by M-CoHOG, I could not report their accuracy for the CALSNIC2 database. Firstly, I estimate the classification accuracy for the widely used ResNet architecture [135] with different depths, such as 10, 18, 50, 101, and 152, and report the best accuracy among them. Secondly, I evaluate the performance of the MobileNet network [151], which utilizes depthwise separable convolutions to develop lightweight deep neural networks. Thirdly, I measure the accuracy with the ShuffleNet framework [141], which employs pointwise group convolution and channel shuffle to provide efficient computation cost. Finally, I calculate the performance with another popular deep model named EfficientNet [138], which effectively balances the network's depth, width, and resolution.

For the 2D CNN-based architectures, I follow similar steps and data, such as input

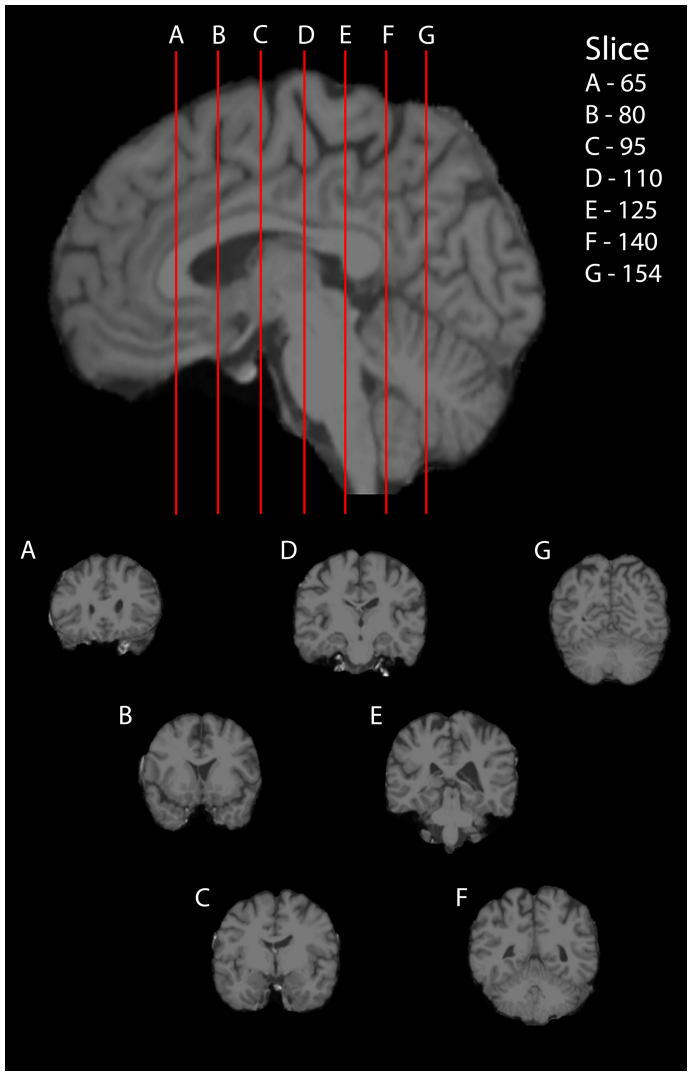


Figure 6.6: Slice number and the corresponding coronal MR image of a T1-weighted scan from an ALS patient. Out of 218 coronal slices, the best performance is found from the slice range of 111 (D) to 125 (E) for the T1-weighted images.

slices and majority voting, as in the proposed method. In contrast, the input for the 3D-based framework is the entire 3D brain MRI to carry out the classification outcome. In Table 6.6, the proposed strategy outperforms all other approaches in most of the evaluation metrics except SPE. The MobileNet architecture achieves higher SPE but produces inferior SEN, indicating a bias towards a particular class.

Additionally, I perform paired t-tests to determine the statistical significance of the five-fold CV ACC of the proposed framework in comparison to other methods. The

Table 6.6: Comparison with popular CNN architectures and previous work for ALS patients vs. healthy controls classification tested on T1-weighted MR images of CALSNIC1 and CALSNIC2 datasets.

Reference	FLOPs	CALSNIC1				CALSNIC2			
	(G)	ACC	SEN	SPE	F1-score	ACC	SEN	SPE	F1-score
ResNet-101 (2D)	23.2	0.717	0.631	0.811	0.699	0.726	0.667	0.783	0.746
ResNet-101 (3D)	243.2	0.720	0.692	0.750	0.720	0.733	0.727	0.739	0.733
ShuffleNet (2D)	2.1	0.680	0.580	0.789	0.654	0.696	0.697	0.696	0.692
ShuffleNet (3D)	14.8	0.680	0.692	0.667	0.692	0.689	0.773	0.609	0.708
MobileNet (2D)	1.0	0.667	0.441	0.911	0.580	0.704	0.546	0.783	0.643
MobileNet (3D)	40.3	0.720	0.539	**0.917**	0.667	0.711	0.546	**0.870**	0.649
EfficientNet (3D)	24.6	0.680	0.667	0.692	0.667	0.711	0.727	0.696	0.696
M-CoHOG (2D)	-	0.745	0.786	0.688	0.752	-	-	-	-
Proposed method	38.4	**0.816**	**0.843**	0.812	**0.815**	**0.818**	**0.824**	0.815	**0.811**

results of these t-tests are reported in Table 6.7, revealing that all p-values are below the significance level of 0.05. This demonstrates that the performance of the proposed method shows statistically significant differences when classifying CALSNIC1 and CALSNIC2 T1-weighted image datasets compared to the other methods.

Table 6.7: The p-values of the paired t-test to compare the statistical significance of the proposed method with other methods for the classification ACC on T1-weighted MR images of CALSNIC1 and CALSNIC2 datasets.

Other methods vs.	p-values	
proposed method	CALSNIC1	CALSNIC2
ResNet-101 (2D)	0.003	0.002
ResNet-101 (3D)	0.004	0.017
ShuffleNet (2D)	0.011	< 0.001
ShuffleNet (3D)	0.022	< 0.001
MobileNet (2D)	0.011	< 0.001
MobileNet (3D)	0.009	0.001
EfficientNet (3D)	0.007	0.003
M-CoHOG (2D)	0.021	-

6.3 Summary

This study presents a comprehensive investigation into the potential of integrating the ViT architecture with spatial and frequency domain features to differentiate ALS

patients from healthy controls. The proposed network outperforms established deep models, achieving superior classification accuracy across T1-weighted, FLAIR, and R2* MRI data. Notably, R2* maps exhibit slightly higher performance among these modalities, highlighting the importance of further exploration to effectively utilize them in ALS diagnosis. Previous attempts at ALS classification have been hindered by small sample sizes or low consistency in accuracy when applied to multi-center data, preventing their translation into clinical diagnosis or trials. The introduced methodology will bring MRI closer to the reality of providing biomarkers for ALS diagnosis and monitoring disease progression as well as response to therapy. In the future, I plan to incorporate clinical features and imaging data to enhance the classification performance. Other neuroimaging modalities, such as functional MRI (fMRI) and diffusion tensor imaging (DTI), can be investigated using a similar framework. The proposed architecture is flexible and can be adapted to other neurodegenerative disease classification tasks with appropriate slice selection, where frequency domain information plays crucial roles in feature extraction.

Chapter 7

Domain Adaptation of MRI Scanners

This chapter introduces a new perspective to address the domain shift issue in multi-site MRI data using unsupervised domain adaptation (UDA) and source-free domain adaptation (SFDA). The primary task is to learn with labelled source samples, and the objective is for the model to function satisfactorily in the target domain without labels. By classifying different scanner manufacturers as distinct domains, the proposed framework can learn better domain-invariant representation and enhance cross-domain classification accuracy. The source code is available at https://github.com/rkushol/DAMS.

7.1 DAMS Method

This study proposes a novel UDA method called DAMS (Domain Adaptation of MRI Scanners) to solve the domain shift issue for MRI data by identifying and addressing the dominant factor causing heterogeneity in the dataset. Unlike previous methods where an entire study/dataset is considered as the source or target domain, the proposed approach demonstrates the necessity of appropriate domain selection for adaptation. Furthermore, the proposed DAMS framework leverages the maximum mean discrepancy (MMD) [39] and a modified deep correlation alignment (CORAL) loss in order to align domain-invariant features.

In a multi-center MRI dataset, domain shift refers to the differences in scanners

and imaging protocols across different sites. Some examples of domain shift parameters include imaging protocol (flip angle, acquisition orientation, slice thickness) and scanner (manufacturer, model, field strength). Therefore, MR images may differ qualitatively and quantitatively from center to center and study to study. Dealing with all of these parameters may appear computationally complex. Interestingly, based on my observations in several MRI datasets, the dominating factor responsible for data deviation is the scanner manufacturer, as shown in Fig 7.1 using the t-SNE [35] method.

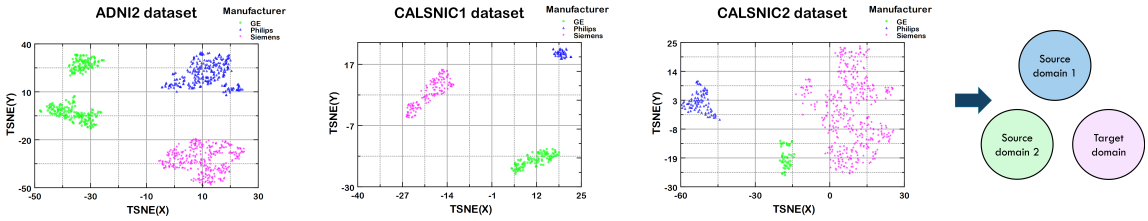


Figure 7.1: Graphs show the distribution of MRI data used in the study from the ADNI [127] and CALSNIC [131] datasets generated by the features of MRQy [57] using t-SNE. Three different colors indicate three different MRI scanner manufacturers' data which are separable from each other. The rightmost panel shows that among three manufacturers, two can be regarded as source domains and the other as the target domain. More findings with different datasets are given in Appendix Fig. B.1.

7.1.1 Problem Formulation

Assume that we have N source domains with labeled samples $\{\mathcal{X}_s^j, \mathcal{Y}_s^j\}_{j=1}^N$, where $\mathcal{X}_s^j$ denotes data from the j^{th} source domain and $\mathcal{Y}_s^j$ are the corresponding class labels. Additionally, we have target domain $\mathcal{X}_t$, with unlabeled $\mathcal{Y}_t$. UDA aims to learn a model that can generalize well to the target domain while minimizing the domain shift between the source and target domains. Specifically, given the source and target domains, the objective is to learn a domain-invariant feature representation $\mathcal{F}$ that can capture the underlying data distributions across different domains. To achieve this, the discrepancy between the source and target feature distributions must be minimized while maintaining the discriminative information necessary for downstream

tasks, such as classification.

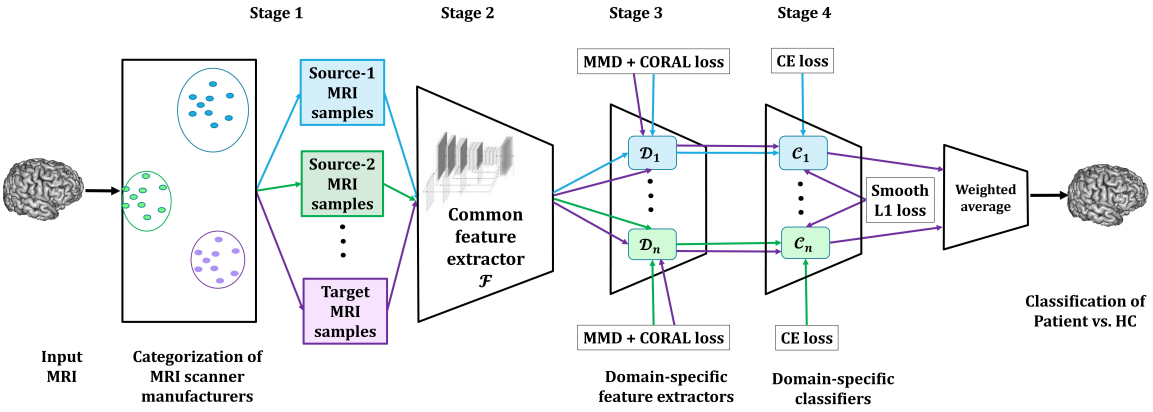


Figure 7.2: The overall workflow of the proposed stages in DAMS framework. Best viewed in color.

7.1.2 Architecture

The overall workflow of the proposed DAMS approach is shown in Fig. 7.2. In stage 1, the proposed framework classifies the domains based on scanner manufacturers (GE, Siemens, Philips). Scanner information is readily available in a standard MRI dataset [29]; however, if this is not available, t-SNE features can be generated using open-source MRQy [57], which can be clustered using K-means clustering to categorize the data based on scanner manufacturers. In stage 2, the common latent feature representation $\mathcal{F}$ is determined from the original feature space of all available domains. Although some methods try to extract domain-invariant features in this shared space, in practice, extracting domain-invariant features across multiple domains leads to a higher degree of discrepancy. Motivated by [39], $\mathcal{F}$ is extended to multiple feature spaces (stage 3), aligning the target domain with available source domains by learning multiple domain-invariant representations $(\mathcal{D}_1, \cdots, \mathcal{D}_n)$ by minimizing the maximum mean discrepancy (MMD) [39] and modified correlation alignment (CORAL) loss. Using these pairwise domain-invariant feature maps, an equal number of domain-specific softmax classifiers $(\mathcal{C}_1, \cdots, \mathcal{C}_n)$ are trained in stage 4, which exploit the Cross-Entropy (CE) loss on the labels of source domains. Finally, to minimize the dissimilarity in

each $(\mathcal{C}_1, \cdots, \mathcal{C}_n)$, the $Smooth_{l1}$ loss function [152] is employed which is less sensitive to outliers. The final target decision (Patient vs. HC) is generated from the weighted average of the outputs of $(\mathcal{C}_1, \cdots, \mathcal{C}_n)$ following [153]. The advantage of having multiple classifiers is that if there are fewer samples from a particular manufacturer, then the data from other manufacturers can be used to achieve better performance.

7.1.3 Loss Functions

To align the feature space of the source and target domains, I leverage the joint contribution from the MMD and deep CORAL loss functions. MMD measures the distance between the empirical mean embeddings of the source and target domains in a reproducing kernel Hilbert space, and the details of $\mathcal{MMD}()$ can be found in [39]. Each feature extractor $(\mathcal{D}_1, \cdots, \mathcal{D}_n)$ learns a domain-invariant map for each pair of source and target domains by minimizing $\mathcal{L}_{mmd}$ as follows:

$$\mathcal{L}_{mmd} = \frac{1}{N} \sum_{j=1}^{N} \mathcal{MMD}(\mathcal{D}_j(\mathcal{F}(X_s^j)), \mathcal{D}_j(\mathcal{F}(X_t))). \tag{7.1}$$

Deep CORAL [40] aims to minimize the difference between the source and target covariance matrices (second-order statistics) in a d-dimensional feature space. I replace the Frobenius norm and normalization term with the mean squared error (MSE) between the covariance matrices of the source (V_s) and target (V_t) feature distributions. The use of MSE provides a more sensitive approach and improved alignment of features which can be defined as follows: $\mathcal{L}_{coral} = MSE(V_s, V_t)$. Each softmax predictor $\mathcal{C}_j$ uses CE classification loss, expressed as follows:

$$\mathcal{L}_{ce} = \sum_{j=1}^{N} \mathbb{E}_{x \sim X_s^j} J(\mathcal{C}_j(\mathcal{D}_j(\mathcal{F}(x_s^j))), y_s^j). \tag{7.2}$$

Since $(\mathcal{C}_1, \cdots, \mathcal{C}_n)$ are trained on diverse source domains, there may be discrepancies in their predictions for target data, in particular, those that are close to decision boundaries. So I employ $Smooth_{l1}$ loss which offers stable gradients for larger values and fewer oscillations during updates to yield a similar classification from each $\mathcal{C}_j$ for

the same target sample. Finally, the total loss is noted as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda(\mathcal{L}_{mmd} + \mathcal{L}_{coral} + \mathcal{L}_{Smooth_{l1}}), \tag{7.3}$$

where λ is the adaptation factor.

7.2 Experiments

7.2.1 Datasets

Publicly available longitudinal datasets from the ADNI [127], AIBL [128] and Minimal Interval Resonance Imaging in Alzheimer's Disease (MIRIAD) [154] are used for classifying AD patients. The CALSNIC [131] multi-center dataset is used for classifying ALS. For ADNI and CALSNIC, two independent versions are used, ADNI1/ADNI2 and CALSNIC1/CALSNIC2, respectively. T1-weighted structural MR images are employed for the above datasets, which are skull-stripped and registered to MNI-152 standard space using the FreeSurfer and FSL software, respectively. After preprocessing, the resulting image dimension is $182 \times 218 \times 182$, and the voxel dimension is converted to 1 mm isotropic resolution. Each dataset's demographics and some scanning protocol details are given in Table 7.1 and Table 3.3, respectively.

7.2.2 Implementation

The proposed DAMS framework employs 32 coronal slices from the central plane of 3D MRI. The final class prediction is performed with the majority voting of these coronal images, similar to [142, 155]. This slice range can better capture significant brain regions related to AD/ALS, including hippocampus, motor cortex, and corticospinal tract. ResNet-50 [135] model is used as the backbone, which is pre-trained on ImageNet [37]. When both training and testing occurred on a specific domain, the data is split as train:validation:test with a 6:2:2 ratio, and mean classification accuracy is reported from five repeated experiments with randomly shuffled data. Avenues of data leakage are avoided following the previous work by Yagis *et al.* [147]. I use the

Table 7.1: Demographic details of the ADNI, AIBL, MIRIAD and CALSNIC datasets

Dataset (#total)	**Group**	**MRI Scanner Manufacturer**					
		GE		**Siemens**		**Philips**	
		Sex (M/F)	**Age** (Mean±Std)	**Sex** (M/F)	**Age** (Mean±Std)	**Sex** (M/F)	**Age** (Mean±Std)
ADNI1	AD	80/80	75.5±7.7	80/80	75.0±7.2	60/49	75.7±7.0
(925)	HC	80/80	75.1±5.7	80/80	75.9±5.9	109/67	75.4±5.2
ADNI2	AD	61/41	75.0±8.5	90/57	75.1±7.8	48/58	74.5±7.3
(852)	HC	64/90	74.3±5.9	92/88	74.0±6.4	69/94	75.6±6.4
AIBL	AD	-	-	28/45	73.6±8.0	-	-
(288)	HC	-	-	99/116	72.9±6.6	-	-
MIRIAD	AD	19/27	69.4±7.1	-	-	-	-
(69)	HC	12/11	69.7±7.2	-	-	-	-
CALSNIC1	ALS	21/25	57.0±11.4	43/28	59.6±10.8	17/1	58.1±9.0
(281)	HC	23/33	50.5±11.9	38/28	57.2±8.1	6/18	53.1±8.4
CALSNIC2	ALS	14/4	54.0±11.8	124/65	60.1±10.2	29/20	62.4±8.2
(546)	HC	18/13	60.1±8.8	120/101	54.9±10.5	12/25	61.7±10.8

SGD optimizer with a momentum of 0.9, a batch size of 32, and $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where $\eta_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$, The learning rate, η_p, during SGD, is modified from 0 to 1 with an iterative scheduling: $\lambda_p = \frac{2}{exp(-\theta p)} - 1$, where $\theta = 10$, to reduce noisy activations in early training phases [156]. The model was trained using an NVIDIA RTX A6000 GPU with 48GB of memory, which took around 8 hours to train on the largest dataset (ADNI1).

7.2.3 Intra-study Validation

While merging a new site into a trained model, the most challenging case is when the scanner manufacturer of the new site differs from those currently involved. To assess the potential impact of such variations in MRI data, I evaluate cross-domain classification accuracy within the ADNI and CALSNIC datasets by considering two scanner manufacturers' data as source domains and the remaining one as the target domain. Table 7.2 shows the results by analyzing three transfer tasks from different domain combinations (source1, source2 → target): GE, Philips → Siemens; GE, Siemens → Philips; Philips, Siemens → GE. The average classification accuracy of AD

vs. HC, when training and testing occur within the same domain, is approximately 90% for both the ADNI1 and ADNI2 datasets but drops to around 80% when the test and the target domains are different (w/o DA). However, by applying the proposed DAMS, the mean classification accuracy returns to about 89%, demonstrating that the proposed method robustly adapts the data from a different scanner to achieve a classification accuracy similar to that of the source domains. The classification accuracy of the CALSNIC datasets (ALS vs. HC) declines by approximately 13% when the test domain differs from the source domains. Nevertheless, an improvement of about 10% in the classification accuracy is obtained using the proposed method. Furthermore, in Table 7.2, an ablation study focusing on the effectiveness of combining the MMD and CORAL loss functions in the network is presented, which slightly outperforms compared to using only one. Finally, the results of two previous multi-source UDA techniques, M3SDA [153] and MFSAN [39] are reproduced, where the proposed DAMS framework surpasses them in terms of classification accuracy.

Table 7.2: The cross-domain intra-study classification accuracy for the ADNI1, ADNI2, CALSNIC1 and CALSNIC2 datasets. [SD: Source Domain, TD: Target Domain, G: GE, S: Siemens, P: Philips]

Dataset	**Training on SD**	**Testing on SD**	**TD**	**Testing on target domain**					
				w/o DA	MMD	CORAL	M3SDA	MFSAN	**DAMS** (Proposed)
ADNI1	G + S	0.90	P	0.80	0.83	0.86	0.82	0.82	**0.88**
	G + P	0.91	S	0.80	0.90	**0.91**	0.85	0.87	**0.91**
	S + P	0.89	G	0.81	**0.87**	0.85	0.84	0.85	**0.87**
ADNI2	G + S	0.89	P	0.79	**0.89**	0.87	0.86	0.87	**0.89**
	G + P	0.92	S	0.81	0.87	0.87	0.84	0.87	**0.88**
	S + P	0.92	G	0.82	0.91	0.90	0.87	0.86	**0.92**
CALS-NIC1	G + S	0.75	P	0.56	0.64	0.62	0.63	0.64	**0.68**
	G + P	0.75	S	0.65	**0.77**	**0.77**	0.72	0.73	**0.77**
	S + P	0.77	G	0.60	0.65	0.66	0.63	0.63	**0.68**
CALS-NIC2	G + S	0.77	P	0.60	0.73	0.72	0.71	0.70	**0.74**
	G + P	0.69	S	0.54	0.59	**0.65**	0.62	0.59	**0.65**
	S + P	0.75	G	0.68	**0.80**	0.76	0.71	0.76	**0.80**

7.2.4 Inter-study Validation

To validate the robustness of the findings, I analyze diverse datasets, including those from single/multiple vendors, as well as datasets containing single/multiple models from the same vendor. Table 7.3 shows the results, and as expected, a noticeable drop in accuracy occurs when the target data differs from the source without DA. However, the proposed DAMS framework substantially improves the accuracy after DA, with accuracy similar to that obtained in the source domains. Another ablation study examines the efficacy of addressing the target domain heterogeneity (TDH). The proposed approach processes the target domain based on the scanner manufacturer, producing better results than considering all samples without TDH. For datasets like AIBL and MIRIAD, the proposed method and the baseline w/o TDH exhibit the same classification accuracy as their data originated from a single manufacturer. Finally, the proposed DAMS method outperforms others in classification accuracy using the same data after domain categorization.

Table 7.3: Inter-study classification accuracy for the ADNI, AIBL, MIRIAD and CALSNIC datasets. [TDH: Target Domain Heterogeneity, SD: Source Domain]

Training on SD	**Testing on SD**	**Target domain**	**Testing on target domain**						
			w/o DA	w/o TDH	MMD	CORAL	M3SDA	MFSAN	**DAMS** (Proposed)
ADNI1	0.89	ADNI2	0.81	0.83	0.88	0.89	0.87	0.87	**0.90**
		AIBL	0.75	**0.84**	0.83	0.82	0.80	0.83	**0.84**
		MIRIAD	0.78	**0.88**	**0.88**	0.85	0.85	**0.88**	**0.88**
ADNI2	0.91	ADNI1	0.79	0.84	0.87	0.87	0.84	0.85	**0.88**
		AIBL	0.74	**0.82**	**0.82**	**0.82**	0.80	0.81	**0.82**
		MIRIAD	0.75	**0.87**	**0.87**	0.85	0.82	**0.87**	**0.87**
CALS-NIC1	0.75	CALS-NIC2	0.61	0.69	0.70	**0.73**	0.68	0.69	**0.73**
CALS-NIC2	0.73	CALS-NIC1	0.64	0.72	0.76	0.74	0.71	0.73	**0.77**

7.2.5 Limitations

The second dominant factor typically observed for data variation is the scanner model. It would be interesting to analyze the results by subdividing each manufacturer's data into different models. The same vendor can reconstruct MRI data differently for various models. However, I have limited my domain consideration to the manufacturer level due to inadequate data availability for different scanner models. Moreover, the proposed DAMS architecture is task-specific and designed for classification. The network's backbone and loss functions require modification of other tasks, such as segmentation and registration.

7.3 BSAMS Method

SFDA is a subfield of domain adaptation where the goal is to adapt a pre-trained model to a new target domain without access to the source domain data. Traditional domain adaptation methods [23–29, 31, 32] typically rely on having access to both the source domain (where the model was initially trained) and target domain data to minimize the distribution discrepancy between them. However, in many practical scenarios, the source domain data might not be available due to privacy concerns, data ownership, or storage constraints. In SFDA, only the trained source model and the target domain data are available. The adaptation process involves adjusting the model parameters or predictions to perform well on the target domain despite the lack of source data.

White-box SFDA provides full access to the pre-trained model, including its internal parameters, intermediate layer outputs, and structure. This access allows for more sophisticated adaptation techniques that can directly manipulate the model. However, through generation techniques like generative adversarial learning [157], it is still possible to recover the raw source data, leaking the individual information. On the other hand, black-box SFDA assumes that only the outputs (predictions) of the

pre-trained model are available for the target domain data. In this scenario, the inner workings of the model (e.g., weights, intermediate representations) are not accessible. Adaptation methods must rely solely on the model's predictions for the target domain data. Therefore, black-box SFDA methods are more appropriate in real-world applications where the labeled data from the source domain might be costly, impractical, or even impossible.

Motivated by the advantages of black-box SFDA, I introduce a novel framework called BSAMS (Black-box Source-free Adaptation of MRI Scanners) to address domain shift issues in multi-center MRI data analysis. Unlike existing SFDA approaches in medical imaging [33, 34, 114–118] that utilize source model parameters during adaptation, the proposed BSAMS method relies solely on the predictions/APIs of the source models, ensuring a more practical and privacy-preserving solution. Furthermore, incorporating self-distillation and consistency regularization enhances the model's robustness into the pseudo label generation and refinement process. Additionally, a learnable ensemble network adaptively integrates information from multiple sources, optimizing the final prediction based on the target domain data. This novel approach substantially improves the accuracy and reliability of MRI data analysis across different centers.

7.3.1 Overview

The overall architecture of the proposed BSAMS framework can be depicted in Fig. 7.3. The goal is to adapt models trained on multiple source domains to a target domain without direct access to the source data during the adaptation process. The proposed BSAMS technique utilizes multiple pre-trained ResNet-50 models and a self-distillation loss for pseudo label generation, followed by consistency regularization for label refinement and a learnable ensemble network for final classification.

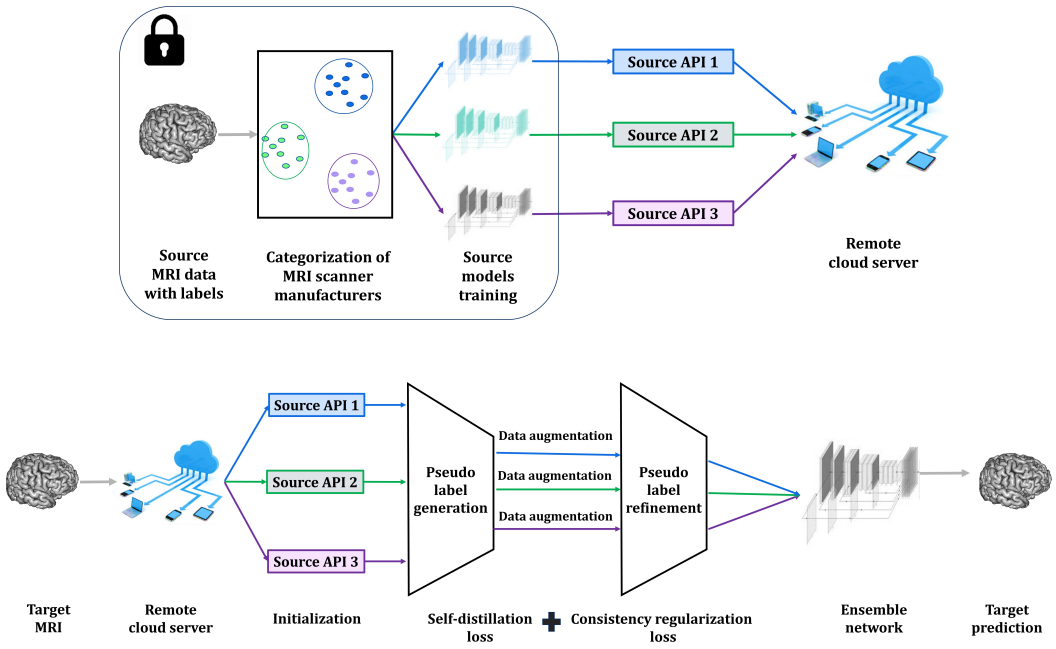


Figure 7.3: The overall architecture of the proposed BSAMS approach. The upper row shows the training process of different source domain data which are categorized based on scanner manufacturer. Three different colors can be assumed data originated from three scanner manufacturers (GE, Philips, Siemens). The trained models are saved in the remote cloud server and only the APIs are available during target domain adaptation. The bottom row highlights the process of target data prediction where pseudo labels are generated using source APIs. Next, the pseudo labels are refined with self-distillation and consistency regularization loss functions. Finally, the target predictions are obtained by utilizing a learnable ensemble network through the refined pseudo labels.

7.3.2 Architecture

Training source domain models

As the large-scale multi-center MRI datasets typically consist of three scanner manufacturers and the domain shift primarily observed due to this factor, I categorize the source domain data considering each scanner manufacturer as a separate domain. ResNet-50 is used as the model architecture, which is pre-trained on the ImageNet dataset. Given source domain datasets $\mathcal{D}_{S1}, \mathcal{D}_{S2}, \mathcal{D}_{S3}$ with labeled samples $(\mathbf{x}_i, y_i)$, three ResNet-50 models f_1, f_2, f_3 are trained using the cross-entropy (CE) loss:

$$\mathcal{L}_{CE}(\mathbf{x}_i, y_i) = -\sum_{c=1}^{C} \mathbf{1}\{y_i = c\} \log p_c(\mathbf{x}_i), \tag{7.4}$$

where $p_c(\mathbf{x}_i)$ is the predicted probability for class c. The trained source models are stored in the remote cloud server and only the predictions are accessible through the source APIs for third-party users.

Generating pseudo labels using self-distillation loss

For the target domain data $\mathcal{D}_T = \{\mathbf{x}_j^T\}$, the trained source models are used to generate soft pseudo labels. The predictions from each model are averaged to produce soft labels:

$$\hat{p}_j = \frac{1}{3} \sum_{k=1}^{3} f_k(\mathbf{x}_j^T). \tag{7.5}$$

The self-distillation loss [109, 158] $\mathcal{L}_{SD}$ encourages the model to output distributions similar to these soft labels:

$$\mathcal{L}_{SD}(\mathbf{x}_j^T, \hat{p}_j) = \mathrm{KL}(f(\mathbf{x}_j^T) \| \hat{p}_j), \tag{7.6}$$

where KL denotes the Kullback-Leibler divergence.

Refining pseudo labels with consistency regularization

To refine pseudo labels, the idea of consistency regularization is employed through data augmentation [159, 160]. The consistency loss $\mathcal{L}_{CR}$ ensures that the model's predictions remain stable across different augmentations $\mathbf{x}_j^{T,\mathrm{aug}}$. I employ random rotation, flipping and adding Gaussian noise to perform augmentation of MRI samples.

$$\mathcal{L}_{CR}(\mathbf{x}_j^T, \mathbf{x}_j^{T,\mathrm{aug}}) = \|f(\mathbf{x}_j^T) - f(\mathbf{x}_j^{T,\mathrm{aug}})\|_2. \tag{7.7}$$

The refined pseudo labels $\tilde{p}_j$ are computed by minimizing both the self-distillation and consistency losses:

$$\mathcal{L}_{total} = \mathcal{L}_{SD} + \lambda \mathcal{L}_{CR}, \tag{7.8}$$

where λ is a weighting factor, empirically set to 0.2 using grid search.

Training the learnable ensemble network

The final step of the proposed approach involves training a learnable ensemble network that aggregates the predictions from the refined pseudo labels produced by the source models. This ensemble network is designed to leverage the strengths of each source model and improve the robustness of the final predictions. The ensemble network concatenates the predictions from the source models and passes them through a fully connected layer to produce the final output. Formally, for a target sample $\mathbf{x}_j^T$:

$$\mathbf{z}_j = [f_1(\mathbf{x}_j^T); f_2(\mathbf{x}_j^T); f_3(\mathbf{x}_j^T)] \tag{7.9}$$

where f_1, f_2, f_3 are the source models and $\mathbf{z}_j$ is the concatenated vector of predictions. This vector is then passed through a fully connected layer with weights $\mathbf{W}$ and $\mathbf{b}$:

$$\hat{y}_j = \text{softmax}(\mathbf{W}\mathbf{z}_j + \mathbf{b}) \tag{7.10}$$

The ensemble model is trained using cross-entropy (CE) loss with the refined pseudo labels $\tilde{p}_j$:

$$\mathcal{L}_{ensemble} = -\sum_{c=1}^{C} \tilde{p}_{jc} \log \hat{y}_{jc}. \tag{7.11}$$

By employing this learnable ensemble network, the diverse strengths of multiple source models are effectively integrated, leading to more accurate and reliable predictions on the target domain.

7.3.3 Results

The proposed BSAMS method is evaluated on the ADNI1, ADNI2, AIBL, and MIRIAD datasets, demonstrating a significant improvement in inter-study AD classification accuracy. This improvement is attributed to the ensemble approach and the effective generation and refinement of pseudo labels. Table 7.4 presents the classification re-

sults, showing a noticeable drop in accuracy when the target data differs from the source. The term without domain adaptation (w/o DA) means a ResNet-50 model is trained with one domain (e.g., ADNI1) and tested with a different domain (e.g., ADNI2). However, the proposed BSAMS framework substantially enhances the accuracy after DA, achieving results similar to or even better than those obtained in the source domains. For instance, training and testing on the ADNI1 dataset yields a classification accuracy of 0.89. Applying BSAMS, the accuracy reaches 0.90 for ADNI1 data, even when using ADNI2 as the source domain. The use of self-distillation (SD) and consistency regularization (CR) loss functions during the pseudo label generation and refinement stages is crucial for achieving higher classification accuracy, as reflected in Table 7.4. The absence of self-distillation (w/o SD) means the network only employs CR loss, and vice versa. The ablation study shows weakness in processing pseudo labels received from different source APIs without both SD and CT loss functions. Additionally, the incorporation of a learnable ensemble network boosts classification performance considerably. Without the ensemble network (w/o ensemble), the classification accuracy declines noticeably, highlighting the effectiveness of the proposed BSAMS framework.

Table 7.4: Inter-study AD classification accuracy for the ADNI, AIBL and MIRIAD datasets. [DA: Domain Adaptation, SD: Self-distillation loss, CR: Consistency Regularization loss]

Training on source domain	**Testing on source domain**	**Target domain**	**Testing on target domain**				
			w/o DA	w/o SD	w/o CR	w/o ensemble	**BSAMS** (Proposed)
ADNI1	0.89	ADNI2	0.81	0.85	0.85	0.87	**0.91**
		AIBL	0.75	0.79	0.80	0.83	**0.85**
		MIRIAD	0.78	0.82	0.85	0.85	**0.87**
ADNI2	0.91	ADNI1	0.79	0.85	0.86	0.87	**0.90**
		AIBL	0.74	0.80	0.82	0.83	**0.84**
		MIRIAD	0.75	0.82	0.82	0.85	**0.87**

7.4 Summary

This study demonstrates the necessity of appropriate domain selection for adaptation instead of considering an entire study/dataset as the source or target domain. The proposed DAMS framework combines MMD and modified CORAL loss functions to extract pairwise domain-specific invariant features. Furthermore, the proposed BSAMS framework eliminates the need for the concurrent access of raw source domain data during target domain adaptation. The issue of scanner bias can negatively impact the reliability of automated analysis of MR images. The proposed solution addresses the undesirable scanner effects of multi-center MRI data and improves the consistency in the classification task of such data. Most importantly, the proposed research not only enables the pooling of data acquired by different sites within a project but also promotes the sharing of data among different studies. The proposed novel strategy can substantially improve the cross-domain classification accuracy of AD/ALS patients from healthy controls.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

MRI datasets play a crucial role in advancing medical research, aiding in the understanding, diagnosis, and treatment of numerous neurological disorders, as well as in training ML models. In recent years, the availability of large-scale multi-center datasets has significantly advanced medical imaging research, which opens the avenues for developing powerful ML algorithms and data-driven methodologies. However, substantial variations from distinct centers, originating from non-biological sources, introduce variability into neuroimaging data, complicating the coherent interpretation of results. ML-based approaches often yield inconsistent outcomes when dealing with data acquired from different MRI scanner models and scanning protocols.

In Chapter 3, I introduce the DSMRI framework, which makes a significant contribution to this field by presenting a simple yet effective unsupervised method for quantifying the degree of domain shift in MRI data. This framework examines a wide range of large multi-center MRI datasets, exploring the impacts of different scanner manufacturers, models, field strengths, and resolutions in the context of domain shift.

Chapter 4 presents another unsupervised framework called DeepDSMRI, which utilizes pre-trained deep models as feature extractors to comprehend the extent of domain shift in MRI datasets. DeepDSMRI proves its efficacy in determining domain shift not only in structural MRI but also in advanced MRI modalities such as fMRI

and DWI. The findings from Chapters 3 and 4 have important implications for advancing the field of medical imaging, enabling more reliable analysis of multi-center MRI datasets. Furthermore, DA and harmonization methods can use these proposed frameworks to validate their effectiveness in reducing or eliminating domain shift.

Chapter 5 comprehensively explores how variations in MRI scanner manufacturers impact different classification tasks using various DL models. The analysis reveals a notable drop in classification accuracy when DL models are tested with data from different scanner manufacturers. Interestingly, the popular statistical approach called ComBat could not improve the performance after harmonizing MRI data across different scanners. Its widespread use could invalidate many published results.

To tackle the challenging task of ALS classification, where current methods fall short, I introduce a novel transformer-based framework named $SF^2Former$. This study pioneers in using a transformer-based deep model for ALS classification, achieving superior performance compared to existing popular DL methods.

Finally, a novel approach to address the domain shift issue in MRI data is developed by addressing the primary factor contributing to heterogeneity within the dataset. A multi-source UDA technique called DAMS is developed to align domain-invariant features between source and target domains by minimizing discrepancies in their feature maps. Instead of treating the entire dataset as a single source or target domain, the method processes the data based on the dominant factor causing variations. Additionally, this research advances the concept of managing domain shift through SFDA, which leverages knowledge from various domains and eliminates the need for concurrent access to source and target data during training. The proposed study is evaluated on diverse neuroimaging datasets, demonstrating a significant improvement in cross-domain disease classification accuracy.

My proposed research directly addresses the challenge of domain shift faced by neuroimaging researchers using multi-center and multi-modal MRI data. This thesis paves the way for future medical image analysis and computer vision communities to

design tools that can more effectively handle extensive and diverse datasets.

8.2 Future Work

Building on the potential demonstrated by the $SF^2Former$ classification framework, several promising directions for future research can be considered:

1. Multi-modal neuroimaging analysis: Developing a fusion ML framework capable of simultaneously analyzing multiple neuroimaging modalities, such as T1-weighted, FLAIR, and DTI, could significantly enhance overall performance. This framework would leverage the unique strengths of each modality to provide a more comprehensive analysis, potentially improving diagnostic accuracy and understanding of neurological conditions.

2. Integration of clinical features: Incorporating clinical features (*e.g.*, UMN/LMN scores, EMG, ALSFRS-R, ECAS) alongside imaging features could further improve the network's prediction accuracy and interpretability. Combining clinical and imaging data can provide a more holistic view of the patient's condition, facilitating more precise and personalized treatment plans.

3. Handling missing modalities: Developing the proposed model to handle missing modalities is crucial for robust performance. This capability would ensure that the framework can still provide accurate predictions even when some imaging or clinical features are absent, making it more applicable in real-world clinical settings where incomplete data is common.

Another promising future investigation could involve implementing existing MRI harmonization techniques and using the proposed domain shift analysis frameworks (*i.e.*, DSMRI and DeepDSMRI) to evaluate their performance before and after harmonization. This could provide valuable insights into the effectiveness of existing harmonization methods in reducing domain shift.

My proposed DA technique considers the scanner manufacturer as the dominant factor of MRI data heterogeneity. However, the scanner model is another significant

source of data variation. Future research could involve subdividing each manufacturer's data into different models to analyze their impacts. Since different models from the same vendor can reconstruct MRI data differently, exploring this aspect could yield deeper insights and further improve the robustness of ML models in handling domain shift.

Bibliography

[1] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico, and F. Sardanelli, "Ai applications to medical images: From machine learning to deep learning," *Physica medica*, vol. 83, pp. 9–24, 2021.

[2] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, 2021.

[3] R. Kushol, M. H. Kabir, M Abdullah-Al-Wadud, and M. S. Islam, "Retinal blood vessel segmentation from fundus image using an efficient multiscale directional representation technique bendlets," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7751–7771, 2020.

[4] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, *et al.*, "Development and validation of an interpretable deep learning framework for alzheimer's disease classification," *Brain*, vol. 143, no. 6, pp. 1920–1933, 2020.

[5] R. Kushol and M. S. Salekin, "Rbvs-net: A robust convolutional neural network for retinal blood vessel segmentation," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 398–402.

[6] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "Synthstrip: Skull-stripping for any brain image," *NeuroImage*, vol. 260, p. 119 474, 2022, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2022.119474.

[7] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *NPJ digital medicine*, vol. 5, no. 1, p. 48, 2022.

[8] S. Kim, E. K. Lee, C. J. Song, and E. Sohn, "Iron rim lesions as a specific and prognostic biomarker of multiple sclerosis: 3t-based susceptibility-weighted imaging," *Diagnostics*, vol. 13, no. 11, p. 1866, 2023.

[9] A. Shoeibi, M. Khodatars, N. Ghassemi, M. Jafari, P. Moridian, R. Alizadehsani, M. Panahiazar, F. Khozeimeh, A. Zare, H. Hosseini-Nejad, *et al.*, "Epileptic seizures detection using deep learning techniques: A review," *International journal of environmental research and public health*, vol. 18, no. 11, p. 5780, 2021.

[10] S. Gull and S. Akbar, "Artificial intelligence in brain tumor detection through mri scans: Advancements and challenges," *Artificial Intelligence and Internet of Things*, pp. 241–276, 2021.

[11] W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao, "The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 623–631.

[12] P. Maggi, P. Sati, G. Nair, I. C. Cortese, S. Jacobson, B. R. Smith, A. Nath, J. Ohayon, V. Van Pesch, G. Perrotta, *et al.*, "Paramagnetic rim lesions are specific to multiple sclerosis: An international multicenter 3t mri study," *Annals of neurology*, vol. 88, no. 5, pp. 1034–1042, 2020.

[13] G. M. E. Elahi, S. Kalra, L. Zinman, A. Genge, L. Korngut, and Y.-H. Yang, "Texture classification of mr images of the brain in als using m-cohog: A multi-center study," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101 659, 2020.

[14] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2713–2724, 2020.

[15] M. Bento, I. Fantini, J. Park, L. Rittner, and R. Frayne, "Deep learning in large and multi-site structural brain mr imaging datasets," *Frontiers in Neuroinformatics*, vol. 15, p. 805 669, 2022.

[16] R. Kushol, A. H. Wilman, S. Kalra, and Y.-H. Yang, "Dsmri: Domain shift analyzer for multi-center mri datasets," *Diagnostics*, vol. 13, no. 18, p. 2947, 2023.

[17] R. Botvinik-Nezer and T. D. Wager, "Reproducibility in neuroimaging analysis: Challenges and solutions," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2022.

[18] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal," *NeuroImage*, vol. 228, p. 117 689, 2021.

[19] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2021.

[20] H. Guan and M. Liu, "Domainatm: Domain adaptation toolbox for medical data analysis," *NeuroImage*, p. 119 863, 2023.

[21] R. K. Gebre, M. L. Senjem, S. Raghavan, C. G. Schwarz, J. L. Gunter, E. I. Hofrenning, R. I. Reid, K. Kantarci, J. Graff-Radford, D. S. Knopman, *et al.*, "Cross–scanner harmonization methods for structural mri may need further work: A comparison study," *NeuroImage*, vol. 269, p. 119 912, 2023.

[22] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-González, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot, *et al.*, "Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation," *Medical image analysis*, vol. 63, p. 101 694, 2020.

[23] C. Wachinger, M. Reuter, and A. D. N. Initiative, "Domain adaptation for alzheimer's disease diagnostics," *Neuroimage*, vol. 139, pp. 470–479, 2016.

[24] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Unlearning scanner bias for mri harmonisation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 369–378.

[25] J. Wolleb, R. Sandkühler, F. Bieder, M. Barakovic, N. Hadjikhani, A. Papadopoulou, Ö. Yaldizli, J. Kuhle, C. Granziera, and P. C. Cattin, "Learn to ignore: Domain adaptation for multi-site mri analysis," in *International Conference on MICCAI*, Springer, 2022, pp. 725–735.

[26] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. d. Leeuw, C. M. Tempany, B. v. Ginneken, *et al.*, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," in *International conference on MICCAI*, Springer, 2017, pp. 516–524.

[27] L.-L. Zeng, Z. Fan, J. Su, M. Gan, L. Peng, H. Shen, and D. Hu, "Gradient matching federated domain adaptation for brain image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[28] A. Ackaouy, N. Courty, E. Vallée, O. Commowick, C. Barillot, and F. Galassi, "Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data," *Frontiers in computational neuroscience*, vol. 14, p. 19, 2020.

[29] R. Wang, P. Chaudhari, and C. Davatzikos, "Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation," *Medical Image Analysis*, vol. 76, p. 102 309, 2022.

[30] R. Kushol, R. Frayne, S. J. Graham, A. H. Wilman, S. Kalra, and Y.-H. Yang, "Domain adaptation of mri scanners as an alternative to mri harmonization," in *MICCAI Workshop on Domain Adaptation and Representation Transfer*, Springer, 2023.

[31] E. Panfilov, A. Tiulpin, S. Klein, M. T. Nieminen, and S. Saarakkala, "Improving robustness of deep learning based knee mri segmentation: Mixup and adversarial domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[32] M. Orbes-Arteaga, T. Varsavsky, C. H. Sudre, Z. Eaton-Rosen, L. J. Haddow, L. Sørensen, M. Nielsen, A. Pai, S. Ourselin, M. Modat, *et al.*, "Multi-domain adaptation in brain mri through paired consistency and adversarial learning," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, Springer, 2019, pp. 54–62.

[33] C. Yang, X. Guo, Z. Chen, and Y. Yuan, "Source free domain adaptation for medical image segmentation with fourier style mining," *Medical Image Analysis*, vol. 79, p. 102 457, 2022.

[34] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed, "Source-free domain adaptation for image segmentation," *Medical Image Analysis*, vol. 82, p. 102 617, 2022.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on CVPR*, 2009, pp. 248–255.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 5989–5996.

[40] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops*, Springer, 2016, pp. 443–450.

[41] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev, "Domain shift in computer vision models for mri data analysis: An overview," in *Thirteenth International Conference on Machine Vision*, SPIE, vol. 11605, 2021, pp. 126–133.

[42] R. Pollitt, "Robustness to domain shifts in mri for deep learning-based methods: A review," 2022.

[43] R. Kushol, P. Parnianpour, A. H. Wilman, S. Kalra, and Y.-H. Yang, "Effects of mri scanner manufacturers in classification tasks with deep learning models," *Scientific Reports*, vol. 13, no. 1, p. 16 791, 2023.

[44] M. Dadar, S. Duchesne, and C. Group, "Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data," *NeuroImage*, vol. 217, p. 116 928, 2020.

[45] D. Tian, Z. Zeng, X. Sun, Q. Tong, H. Li, H. He, J.-H. Gao, Y. He, and M. Xia, "A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset," *NeuroImage*, vol. 257, p. 119 297, 2022.

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

[47] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu, "Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects," *arXiv preprint arXiv:1910.04597*, 2019.

[48] 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*, vol. 13, p. 729, 2019.

[49] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, "Mriqc: Advancing the automatic prediction of image quality in mri from unseen sites," *PloS one*, vol. 12, no. 9, e0184661, 2017.

[50] A. Keshavan, E. Datta, I. M. McDonough, C. R. Madan, K. Jordan, and R. G. Henry, "Mindcontrol: A web application for brain segmentation quality control," *NeuroImage*, vol. 170, pp. 365–372, 2018.

[51] M. E. Osadebey, M. Pedersen, D. L. Arnold, K. E. Wendel-Mitoraj, and f. t. Alzheimer's Disease Neuroimaging Initiative, "Standardized quality metric system for structural brain magnetic resonance images in multi-center neuroimaging study," *BMC medical imaging*, vol. 18, pp. 1–19, 2018.

[52] J. Jang, K. Bang, H. Jang, D. Hwang, and A. D. N. Initiative, "Quality evaluation of no-reference mr images using multidirectional filters and image statistics," *Magnetic resonance in medicine*, vol. 80, no. 3, pp. 914–924, 2018.

[53] O. Esteban, R. W. Blair, D. M. Nielson, J. C. Varada, S. Marrett, A. G. Thomas, R. A. Poldrack, and K. J. Gorgolewski, "Crowdsourced mri quality metrics and expert quality annotations for training of humans and machines," *Scientific data*, vol. 6, no. 1, p. 30, 2019.

[54] M. Oszust, A. Piórkowski, and R. Obuchowicz, "No-reference image quality assessment of magnetic resonance images with high-boost filtering and local features," *Magnetic Resonance in Medicine*, vol. 84, no. 3, pp. 1648–1660, 2020.

[55] S. Bottani, N. Burgos, A. Maire, A. Wild, S. Ströer, D. Dormont, O. Colliot, and A. S. Group, "Automatic quality control of brain t1-weighted magnetic resonance images for a clinical data warehouse," *Medical Image Analysis*, vol. 75, p. 102 219, 2022.

[56] I. Stępień and M. Oszust, "A brief survey on no-reference image quality assessment methods for magnetic resonance images," *Journal of Imaging*, vol. 8, no. 6, p. 160, 2022.

[57] A. R. Sadri, A. Janowczyk, R. Zhou, R. Verma, N. Beig, J. Antunes, A. Madabhushi, P. Tiwari, and S. E. Viswanath, "Mrqy—an open-source tool for quality control of mr imaging data," *Medical physics*, vol. 47, no. 12, pp. 6029–6038, 2020.

[58] L.-H. Kuan, P. Parnianpour, R. Kushol, N. Kumar, T. Anand, S. Kalra, and R. Greiner, "Accurate personalized survival prediction for amyotrophic lateral sclerosis patients," *Scientific Reports*, vol. 13, no. 1, p. 20 713, 2023.

[59] J. P. Taylor, R. H. Brown Jr, and D. W. Cleveland, "Decoding als: From genes to mechanism," *Nature*, vol. 539, no. 7628, pp. 197–206, 2016.

[60] M. K. Jaiswal, "Riluzole and edaravone: A tale of two amyotrophic lateral sclerosis drugs," *Medicinal Research Reviews*, vol. 39, no. 2, pp. 733–748, 2019.

[61] C. A. Sage, R. R. Peeters, A. Görner, W. Robberecht, and S. Sunaert, "Quantitative diffusion tensor imaging in amyotrophic lateral sclerosis," *Neuroimage*, vol. 34, no. 2, pp. 486–499, 2007.

[62] R. Maani, Y.-H. Yang, D. Emery, and S. Kalra, "Cerebral degeneration in amyotrophic lateral sclerosis revealed by 3-dimensional texture analysis," *Frontiers in neuroscience*, vol. 10, p. 120, 2016.

[63] A. Ignjatović, Z. Stević, S. Lavrnić, M. Daković, and G. Bačić, "Brain iron mri: A biomarker for amyotrophic lateral sclerosis," *Journal of magnetic resonance imaging*, vol. 38, no. 6, pp. 1472–1479, 2013.

[64] C. Wang, S. Foxley, O. Ansorge, S. Bangerter-Christensen, M. Chiew, A. Leonte, R. A. Menke, J. Mollink, M. Pallebage-Gamarallage, M. R. Turner, *et al.*, "Methods for quantitative susceptibility and r2* mapping in whole post-mortem brains at 7t applied to amyotrophic lateral sclerosis," *NeuroImage*, vol. 222, p. 117 216, 2020.

[65] M. Hecht, F Fellner, C Fellner, M. Hilz, D Heuss, and B Neundörfer, "Mri-flair images of the head show corticospinal tract alterations in als patients more frequently than t2-, t1-and proton-density-weighted images," *Journal of the neurological sciences*, vol. 186, no. 1-2, pp. 37–44, 2001.

[66] J. Jin, F. Hu, Q. Zhang, R. Jia, and J. Dang, "Hyperintensity of the corticospinal tract on flair: A simple and sensitive objective upper motor neuron degeneration marker in clinically verified amyotrophic lateral sclerosis," *Journal of the neurological sciences*, vol. 367, pp. 177–183, 2016.

[67] J. Fabes, L. Matthews, N. Filippini, K. Talbot, M. Jenkinson, and M. R. Turner, "Quantitative flair mri in amyotrophic lateral sclerosis," *Academic radiology*, vol. 24, no. 10, pp. 1187–1194, 2017.

[68] H. Li, Q. Zhang, Q. Duan, J. Jin, F. Hu, J. Dang, and M. Zhang, "Brainstem involvement in amyotrophic lateral sclerosis: A combined structural and diffusion tensor mri analysis," *Frontiers in Neuroscience*, vol. 15, p. 675 444, 2021.

[69] L. C. Alberich, J. F. Vázquez-Costa, A. Ten-Esteve, M. Mazón, and L. Martí-Bonmatí, "Imaging biomarkers in amyotrophic lateral sclerosis," *Neurodegenerative Diseases Biomarkers: Towards Translating Research to Clinical Practice*, pp. 507–548, 2022.

[70] X. Liu, F. Xing, C. Yang, C.-C. J. Kuo, S. Babu, G. El Fakhri, T. Jenkins, and J. Woo, "Voxelhop: Successive subspace learning for als disease classification using structural mri," *IEEE journal of biomedical and health informatics*, vol. 26, no. 3, pp. 1128–1139, 2021.

[71] J. Thome, R. Steinbach, J. Grosskreutz, D. Durstewitz, and G. Koppe, "Classification of amyotrophic lateral sclerosis by brain volume, connectivity, and network dynamics," *Human brain mapping*, vol. 43, no. 2, pp. 681–699, 2022.

[72] Q.-F. Chen, X.-H. Zhang, N.-X. Huang, and H.-J. Chen, "Identification of amyotrophic lateral sclerosis based on diffusion tensor imaging and support vector machine," *Frontiers in neurology*, vol. 11, p. 275, 2020.

[73] T. D. Kocar, A. Behler, A. C. Ludolph, H.-P. Müller, and J. Kassubek, "Multiparametric microstructural mri and machine learning classification yields high diagnostic accuracy in amyotrophic lateral sclerosis: Proof of concept," *Frontiers in neurology*, vol. 12, 2021.

[74] M. A. Ebrahimighahnavieh, S. Luo, and R. Chiong, "Deep learning to detect alzheimer's disease from neuroimaging: A systematic literature review," *Computer methods and programs in biomedicine*, vol. 187, p. 105 242, 2020.

[75] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C. E. Teunissen, J. Cummings, and W. M. van der Flier, "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021.

[76] K. Gunawardena, R. Rajapakse, and N. Kodikara, "Applying convolutional neural networks for pre-detection of alzheimer's disease from structural mri data," in *International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, 2017, pp. 1–7.

[77] H. Shahamat and M. S. Abadeh, "Brain mri analysis using a deep learning based evolutionary approach," *Neural Networks*, vol. 126, pp. 218–234, 2020.

[78] B. Solano-Rojas, R. Villalón-Fonseca, and G. Marín-Raventós, "Alzheimer's disease early detection using a low cost three-dimensional densenet-121 architecture," in *International Conference on Smart Homes and Health Telematics*, 2020, pp. 3–15.

[79] A. Ebrahimi, S. Luo, R. Chiong, and A. D. N. Initiative, "Deep sequence modelling for alzheimer's disease detection using mri," *Computers in Biology and Medicine*, p. 104 537, 2021.

[80] Y. Zhang, Q. Teng, Y. Liu, Y. Liu, and X. He, "Diagnosis of alzheimer's disease based on regional attention with smri gray matter slices," *Journal of neuroscience methods*, vol. 365, p. 109 376, 2022.

[81] W. Lin, T. Tong, Q. Gao, D. Guo, X. Du, Y. Yang, G. Guo, M. Xiao, M. Du, X. Qu, *et al.*, "Convolutional neural networks-based mri image analysis for the alzheimer's disease prediction from mild cognitive impairment," *Frontiers in neuroscience*, vol. 12, p. 777, 2018.

[82] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[83] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, and A. D. N. Initiative, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease," *Neuroimage*, vol. 208, p. 116 459, 2020.

[84] F. Li, M. Liu, and A. D. N. Initiative, "Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks," *Computerized Medical Imaging and Graphics*, vol. 70, pp. 101–110, 2018.

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[86] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *International Conference on Computer Vision (ICCV)*, 2021.

[87] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[88] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, PMLR, 2021, pp. 10 347–10 357.

[89] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[90] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 14–24.

[91] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 36–46.

[92] T. C. W. Mok and A. C. S. Chung, "Affine medical image registration with coarse-to-fine vision transformer," *arXiv preprint arXiv:2203.15216v2*, 2022.

[93] T. Stegmüller, A. Spahr, B. Bozorgtabar, and J.-P. Thiran, "Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification," *arXiv preprint arXiv:2202.07570*, 2022.

[94] Z. Cai, H. He, L. Lin, and X. Tang, "Uni4eye: Unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification," *arXiv preprint arXiv:2203.04614*, 2022.

[95] J. Cheng, A. V. Dalca, B. Fischl, L. Zöllei, and A. D. N. Initiative, "Cortical surface registration using unsupervised learning," *NeuroImage*, vol. 221, p. 117 161, 2020.

[96] G. Modanwal, A. Vellal, M. Buda, and M. A. Mazurowski, "Mri image harmonization using cycle-consistent generative adversarial network," in *Medical Imaging 2020: Computer-Aided Diagnosis*, SPIE, vol. 11314, 2020, pp. 259–264.

[97] M. Liu, P. Maiti, S. Thomopoulos, A. Zhu, Y. Chai, H. Kim, and N. Jahanshad, "Style transfer using generative adversarial networks for multi-site mri harmonization," in *International Conference on MICCAI*, Springer, 2021, pp. 313–322.

[98] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, pp. 104–120, 2018.

[99] J. Radua, E. Vieta, R. Shinohara, P. Kochunov, Y. Quidé, M. J. Green, C. S. Weickert, T. Weickert, J. Bruggemann, T. Kircher, *et al.*, "Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma," *NeuroImage*, vol. 218, p. 116 956, 2020.

[100] N. Maikusa, Y. Zhu, A. Uematsu, A. Yamashita, K. Saotome, N. Okada, K. Kasai, K. Okanoya, O. Yamashita, S. C. Tanaka, *et al.*, "Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics," *Human brain mapping*, vol. 42, no. 16, pp. 5278–5287, 2021.

[101] T. Itahashi, Y. Y. Aoki, A. Yamashita, T. Soda, J. Fujino, H. Ohta, R. Aoki, M. Nakamura, N. Kato, S. C. Tanaka, *et al.*, "Effects of upgrading acquisition-techniques and harmonization methods: A multi-modal mri study with implications for longitudinal designs," *bioRxiv*, 2021.

[102] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, *et al.*, "Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, p. 116 450, 2020.

[103] H. Horng, A. Singh, B. Yousefi, E. A. Cohen, B. Haghighi, S. Katz, P. B. Noël, R. T. Shinohara, and D. Kontos, "Generalized combat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects," *Scientific reports*, vol. 12, no. 1, pp. 1–12, 2022.

[104] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, "Advancing medical imaging informatics by deep learning-based domain adaptation," *Yearbook of medical informatics*, vol. 29, no. 01, pp. 129–138, 2020.

[105] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *Computers in Biology and Medicine*, p. 107 912, 2023.

[106] M. Ye, J. Zhang, J. Ouyang, and D. Yuan, "Source data-free unsupervised domain adaptation for semantic segmentation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 2233–2242.

[107] Z. Qiu, Y. Zhang, H. Lin, S. Niu, Y. Liu, Q. Du, and M. Tan, "Source-free domain adaptation via avatar prototype generation and adaptation," *arXiv preprint arXiv:2106.15326*, 2021.

[108] M. Jing, J. Li, K. Lu, L. Zhu, and H. T. Shen, "Visually source-free domain adaptation via adversarial style matching," *IEEE Transactions on Image Processing*, 2024.

[109] J. Liang, D. Hu, J. Feng, and R. He, "Dine: Domain adaptation from single and multiple black-box predictors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8003–8013.

[110] J. Yang, X. Peng, K. Wang, Z. Zhu, J. Feng, L. Xie, and Y. You, "Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors," *arXiv preprint arXiv:2205.14467*, 2022.

[111] J. Zhang, J. Huang, X. Jiang, and S. Lu, "Black-box unsupervised domain adaptation with bi-directional atkinson-shiffrin memory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 771–11 782.

[112] Y. Fang, P.-T. Yap, W. Lin, H. Zhu, and M. Liu, "Source-free unsupervised domain adaptation: A survey," *arXiv preprint arXiv:2301.00265*, 2022.

[113] Z. Yu, J. Li, Z. Du, L. Zhu, and H. T. Shen, "A comprehensive survey on source-free domain adaptation," *arXiv preprint arXiv:2302.11803*, 2023.

[114] J. Hong, Y.-D. Zhang, and W. Chen, "Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation," *Knowledge-Based Systems*, vol. 250, p. 109 155, 2022.

[115] S. Kondo, "Source-free unsupervised domain adaptation with norm and shape constraints for medical image segmentation," *arXiv preprint arXiv:2209.01300*, 2022.

[116] Y. Wang, J. Cheng, Y. Chen, S. Shao, L. Zhu, Z. Wu, T. Liu, and H. Zhu, "Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation," *arXiv preprint arXiv:2304.13672*, 2023.

[117] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Sfharmony: Source free domain adaptation for distributed neuroimaging analysis," *arXiv preprint arXiv:2303.15965*, 2023.

[118] Q. Yu, N. Xi, J. Yuan, Z. Zhou, K. Dang, and X. Ding, "Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning," *arXiv preprint arXiv:2307.09769*, 2023.

[119] Y. Wang, Y. Zhang, W. Xuan, E. Kao, P. Cao, B. Tian, K. Ordovas, D. Saloner, and J. Liu, "Fully automatic segmentation of 4d mri for cardiac functional measurements," *Medical physics*, vol. 46, no. 1, pp. 180–189, 2019.

[120] V. A. Magnotta, L. Friedman, and F. BIRN, "Measurement of signal-to-noise and contrast-to-noise in the fbirn multicenter imaging study," *Journal of digital imaging*, vol. 19, pp. 140–147, 2006.

[121] C. Hui, Y. X. Zhou, and P. Narayana, "Fast algorithm for calculation of inhomogeneity gradient in magnetic resonance imaging data," *Journal of Magnetic Resonance Imaging*, vol. 32, no. 5, pp. 1197–1208, 2010.

[122] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[123] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "Pywavelets: A python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, no. 36, p. 1237, 2019.

[124] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[125] D. Gadkari, "Image quality analysis using glcm," 2004.

[126] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in python," *PeerJ*, vol. 2, e453, 2014.

[127] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[128] K. A. Ellis, A. I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N. T. Lautenschlager, N. Lenzo, R. N. Martins, P. Maruff, *et al.*, "The australian imaging, biomarkers and lifestyle (aibl) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease," *International psychogeriatrics*, vol. 21, no. 4, pp. 672–687, 2009.

[129] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[130] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, *et al.*, "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[131] S. Kalra, M. Khan, L. Barlow, C. Beaulieu, M. Benatar, H. Briemberg, S. Chenji, M. G. Clua, S. Das, A. Dionne, *et al.*, "The canadian als neuroimaging consortium (calsnic)-a multicentre platform for standardized imaging and clinical studies in als," *MedRxiv*, pp. 2020–07, 2020.

[132] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[133] R. Kushol, S. Kalra, and Y.-H. Yang, "Deepdsmri: Deep domain shift analyzer for mri," in *Annual Conference on Medical Image Understanding and Analysis*, Springer, 2024, pp. 81–95.

[134] R. Wightman, *Pytorch image models*, https://github.com/rwightman/pytorch-image-models, 2019. DOI: 10.5281/zenodo.4414861.

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

[136] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[137] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[138] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.

[139] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9981–9990.

[140] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *International Conference on Machine Learning*, PMLR, 2021, pp. 2286–2296.

[141] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[142] R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, and Y.-H. Yang, "Addformer: Alzheimer's disease detection from structural mri using fusion transformer," in *IEEE 19th International Symposium on Biomedical Imaging*, IEEE, 2022, pp. 1–5.

[143] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[144] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[145] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, *et al.*, "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[146] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[147] E. Yagis, S. W. Atnafu, A. García, C. Marzi, R. Scheda, M. Giannelli, C. Tessa, L. Citi, and S. Diciotti, "Effect of data leakage in brain mri classification using 2d convolutional neural networks," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[148] W. Zhu, L. Sun, J. Huang, L. Han, and D. Zhang, "Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri," *IEEE Transactions on Medical Imaging*, 2021.

[149] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[150] W. Yan, L. Huang, L. Xia, S. Gu, F. Yan, Y. Wang, and Q. Tao, "Mri manufacturer shift and adaptation: Increasing the generalizability of deep learning segmentation for mr images acquired with different scanners," *Radiology: Artificial Intelligence*, vol. 2, no. 4, 2020.

[151] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[152] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[153] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.

[154] I. B. Malone, D. Cash, G. R. Ridgway, D. G. MacManus, S. Ourselin, N. C. Fox, and J. M. Schott, "Miriad—public release of a multiple time point alzheimer's mr imaging dataset," *NeuroImage*, vol. 70, pp. 33–36, 2013.

[155] R. Kushol, C. C. Luk, A. Dey, M. Benatar, H. Briemberg, A. Dionne, N. Dupré, R. Frayne, A. Genge, S. Gibson, *et al.*, "Sf2former: Amyotrophic lateral sclerosis identification from multi-center mri data using spatial and frequency fusion transformer," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102 279, 2023.

[156] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.

[157] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[158] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, 2021.

[159] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3733–3748, 2021.

[160] S. Yang, Y. Dong, R. Ward, I. S. Dhillon, S. Sanghavi, and Q. Lei, "Sample efficiency of data augmentation consistency regularization," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2023, pp. 3825–3853.

Appendix A: First Appendix

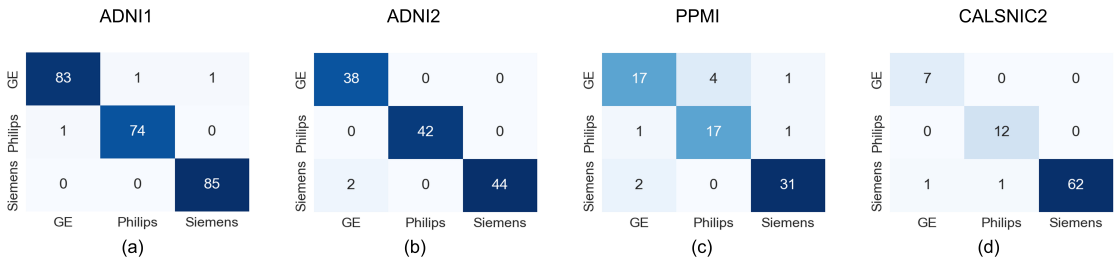


Figure A.1: MRI scanner manufacturer classification results for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets generated by ShuffleNetV2 model. The classification accuracy is approximately 99% for the (a) ADNI1, (b) ADNI2, and (d) CALSNIC2 datasets whereas the accuracy is around 94% for the (c) PPMI dataset.

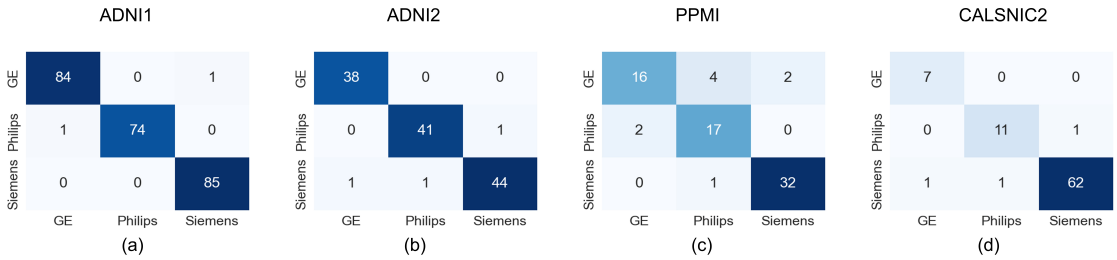


Figure A.2: MRI scanner manufacturer classification results for the ADNI1, ADNI2, PPMI, and CALSNIC2 datasets generated by MobileNetV2 model. The classification accuracy is approximately 98% for the (a) ADNI1, (b) ADNI2, and (d) CALSNIC2 datasets whereas the accuracy is around 94% for the (c) PPMI dataset.

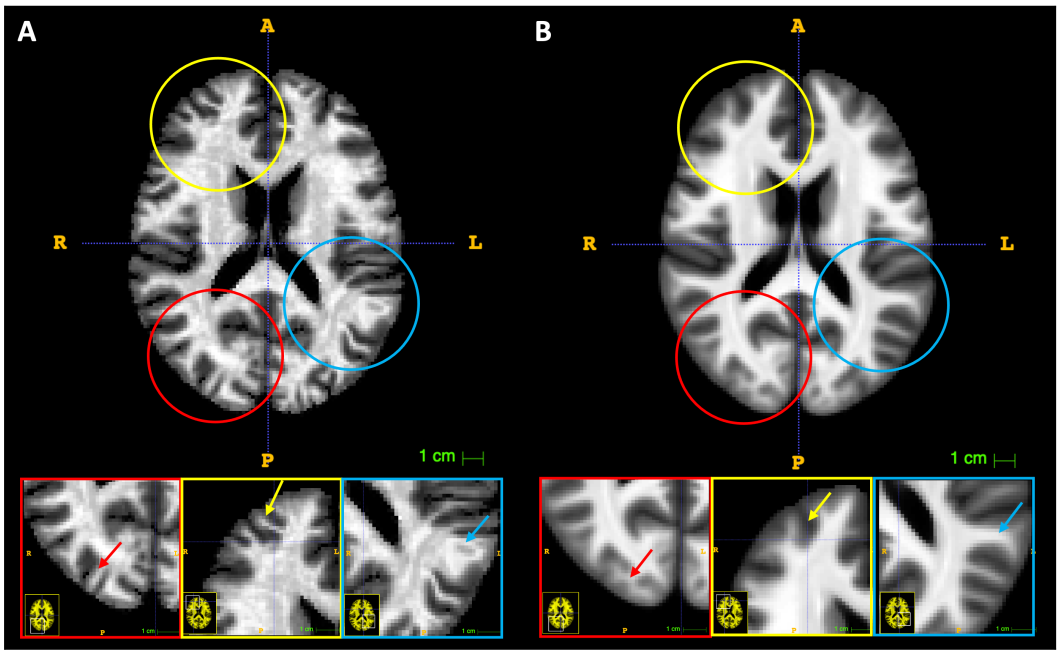
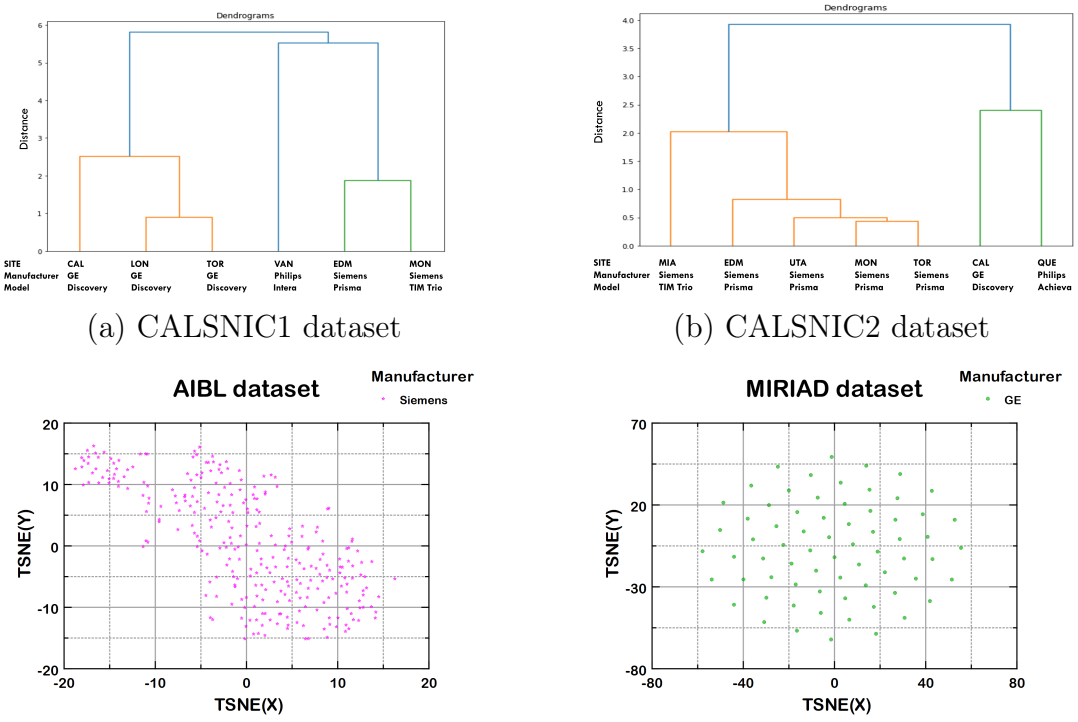


Figure A.3: Undesirable effects in voxel-wise ComBat-GAM harmonization of structural MRI. A) One 2D axial slice of 3D MR image of CALSNIC2 dataset before harmonization, B) corresponding slice after harmonization. The red, yellow, and blue arrows point to the regions with manipulated structures, including the disappearance of details or abnormal shape changes, resulting from the ComBat-GAM harmonization.

Appendix B: Second Appendix



(a) CALSNIC1 dataset

(b) CALSNIC2 dataset

(c) AIBL dataset

(d) MIRIAD dataset

Figure B.1: Panel (a) illustrates the hierarchical clustering dendrogram for six sites comprising the CALSNIC1 dataset, where the features are generated from the average of evaluation metrics measured in MRQy. In (b), similar findings are observed for seven sites of the CALSNIC2 dataset. The MRI manufacturer is the primary factor in grouping site effects into clusters, followed by the scanner model. The t-SNE plots for the AIBL and MIRIAD datasets are shown in (c) and (d). Both AIBL and MIRIAD consist of MRI data from a single scanner manufacturer. Therefore, no significant clustering is noticeable in their data distribution.