Enhancing Alzheimer's Prognostic Models with Cross-Domain Self-Supervised Learning and MRI Data Harmonization

by

Saba Dadsetan

Bachelor of Science, University of Tehran, 2017

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

INTELLIGENT SYSTEMS PROGRAM

This dissertation was presented

by

Saba Dadsetan

It was defended on

July 2, 2024

and approved by

Prof. Dana L. Tudorascu, Department of Psychiatry and Biostatistics

Prof. Ahmad P. Tafti, Department of Health Information Management

Prof. Yalini Senathirajah, Department of Biomedical Informatics

Prof. Davneet S. Minhas, Department of Radiology

Abstract

Enhancing Alzheimer's Prognostic Models with Cross-Domain Self-Supervised Learning and MRI Data Harmonization

Saba Dadsetan, PhD

University of Pittsburgh, 2024

In the rapidly evolving field of medical imaging, the development of effective artificial intelligence systems requires both advanced deep learning algorithms and substantial, high-quality datasets. However, the acquisition and annotation of such data, particularly in specialized domains like clinical disease prognostics, is often prohibitively expensive and time-consuming. This research explores the potential of cross-domain self-supervised learning (CDSSL) as an innovative solution to these challenges, with a specific focus on enhancing Alzheimer's disease progression models using brain Magnetic Resonance Imaging (MRI) data.

Our study introduces a novel CDSSL approach tailored for disease prognostic modeling, emphasizing regression tasks in medical imaging. Using Alzheimer's disease progression prediction from brain MRI as a case study, we demonstrate that self-supervised pretraining significantly improves prognostic accuracy. Notably, models pretrained on extended, unlabeled brain MRI datasets consistently outperform those using natural images, with an optimal combination of both data sources yielding the best results.

Furthermore, we address the critical issue of data harmonization in medical imaging, investigating the impact of scanner-specific variations arising from diverse manufacturers and models. Our findings highlight CDSSL's potential in ensuring data consistency across different scanner environments, thereby enhancing data comparability and reproducibility. Specifically, we propose two methods Augmentation CDSSL and Auxiliary CDSSL, and show improved prognostic model and scanner variability reduction.

Additionally, we compare our methods with an unsupervised harmonization model, demonstrating that our approach achieves better results in most of the datasets. This research underscores the significance of scanner-aware self-supervised learning in refining

medical imaging methodologies, particularly in the context of Alzheimer's disease (AD) progression modeling. The proposed approach not only improves model accuracy and robustness in limited data scenarios but also offers a promising solution for mitigating scanner variability. These advancements have profound implications for the application of Artificial Intelligence (AI) in clinical settings, potentially leading to more accurate and reliable prognostic tools for AD.

Table of Contents

List of Tables

List of Figures

Preface

I would like to dedicate this work to all the courageous women worldwide who are fighting for their lives and freedom.

Also, I am deeply grateful to my advisor, Prof. Dana L. Tudorascu, whose unwavering support, encouragement, patience, and invaluable guidance have been a cornerstone of my Ph.D. journey. It has been a profound privilege to work under the mentorship of such a wise, attentive, and knowledgeable professor.

I extend my heartfelt appreciation to my dissertation committee members, Prof. Ahmad P. Tafti, Prof. Yalini Senathirajah, and Prof. Davneet S. Minhas, for their generous commitment of time and their insightful feedback, which has significantly enriched this work.

My thanks also go to my friends and colleagues, whose support has been invaluable throughout my graduate studies. I am especially thankful to my parents, Mahin and Samad, my brother, Sina, and my in-laws for their unwavering encouragement. A special note of gratitude goes to my dearest best friend and husband, Alireza, whose unconditional love and support have been my anchor during the most challenging times.

Acknowledgement

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

Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following organizations: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company

Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

The Australian Imaging Biomarkers and Lifestyle (AIBL) study (www.AIBL.csiro.au) is a consortium between Austin Health, CSIRO, Edith Cowan University, the Florey Institute (The University of Melbourne), and the National Aging Research Institute. Partial financial support was provided by the Alzheimer's Association (US), the Alzheimer's Drug Discovery Foundation, an anonymous foundation, the Science and Industry Endowment Fund, the Dementia Collaborative Research Centres, the Victorian Government's Operational Infrastructure Support program, the McCusker Alzheimer's Research Foundation, the National Health and Medical Research Council, and the Yulgilbar Foundation. Numerous commercial interactions have supported data collection and analysis. In-kind support has also been provided by Sir Charles Gairdner Hospital, Cogstate Ltd., Hollywood Private Hospital, the University of Melbourne, and St. Vincent's Hospital.

Data were provided in part by OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

The Harvard Aging Brain Study (HABS - P01AG036694; https://habs.mgh.harvard.edu) was launched in 2010, funded by the National Institute on Aging. and is led by principal investigators Reisa A. Sperling MD and Keith A. Johnson MD at Massachusetts General

Hospital/Harvard Medical School in Boston, MA.

The Wisconsin Registry for Alzheimer's Prevention (WRAP) dataset were supported with grants from the US National Institutes of Health (grant Nos. AG027161 and AG021155).

1.0 Introduction

The scanner effect, often referred to as the "batch effect" in medical imaging, describes variability introduced in medical images due to differences in scanning devices, protocols, and settings. Such variability can hide real biological differences or clinical findings and introduce artificial discrepancies that may hinder accurate analysis (O'Brien et al., 2016). The scanner effect poses several challenges. One of the most significant challenges is in multi-center studies, where imaging data is collected from different devices with varying protocols. Inconsistent data can lead to reduced statistical power or inaccurate findings. Such variations can be especially problematic when machine learning models are used, as they may become adept at identifying scanner-specific patterns rather than true clinical indicators (Zhang et al., 2018). The sources of scanner variability can be multifaceted. They can arise from (a) Different manufacturers or models of scanners (Smith et al., 2017). (b) Variations in imaging protocols, such as different MRI pulse sequences or CT radiation doses. (c) Calibration and maintenance differences over time. Patient positioning and physiological conditions (e.g., heart rate) during scanning (Nyúl et al., 2000).

Several techniques also called "harmonization", have been developed to reduce or account for the scanner effect. Harmonization techniques aim to make images from different scanners or protocols more comparable. These techniques include: 1- Image normalization and standardization (Fortin et al., 2016; Shinohara et al., 2014). 2- ComBat, a method adapted from genomics to adjust for batch effects in imaging (Fortin et al., 2017). 3- Supervised and unsupervised machine learning-based approaches, designed to identify and mitigate scanner-specific patterns (Tustison et al., 2018).

While harmonization techniques have shown promise, they are not without challenges. In some cases, harmonization can overcorrect or introduce new artificial patterns into the data. Furthermore, there's the risk of potentially removing clinically relevant information embedded in the images (Pomponio et al., 2020).

Given the limitations of existing methods, there's a growing interest in exploring innovative approaches such as self-supervised learning. Self-supervised learning models are trained

using data as their supervision and are emerging as a potential solution. This offers the advantage of learning meaningful representations without extensive labeled data, potentially capturing and addressing scanner variability (Zhou et al., 2020).

1.1 Motivation

The field of medical image analysis has recently been confronted with the challenge of the "scanner effect." By advancing imaging acquisition technologies, this phenomenon, in which different scanning devices or protocols produce varied results, can significantly hinder the consistency and reliability of medical image analysis. The implications are profound: errors or inconsistencies could lead to misdiagnoses or ineffective treatment planning. The issue is further accentuated when we consider diseases like Alzheimer's disease, where early and accurate detection can pave the way for better patient outcomes.

Yet, the scanner effect is not merely a challenge—it's an opportunity. In this thesis, we proposed solutions to this problem, which can potentially revolutionize how medical imaging data is interpreted using better consistency, reliability, and generalizability across devices and protocols.

1.2 Thesis Statement

This thesis delved into the potential of Self-Supervised Learning (SSL) techniques, specifically when applied cross-domain (spanning both natural and medical images), as a means to address and mitigate the scanner effect. Within this exploration, the following hypotheses are proposed:

- **H1. Cross-Domain Enhancement**: Leveraging cross-domain SSL techniques (i.e. utilizing both natural and medical images) will significantly outperform domain-specific self-supervised techniques in mitigating the scanner effect when applied to supervised tasks.

- **H2. Scanner-Aware SSL Optimization**: Incorporating scanner-specific information during the SSL process will not only reduce the scanner effect more effectively but will also preserve and potentially enhance clinically relevant information within the images.
- **H3. SSL Efficacy Over Traditional Harmonization**: SSL techniques, when appropriately employed, can achieve a more substantial reduction in scanner effects in medical imaging when compared to traditional unsupervised harmonization techniques.

Within the scope of this thesis, we investigate the merits and drawbacks of existing harmonization techniques and explore innovative ways to incorporate scanner-specific information during SSL, aiming to validate these hypotheses and further understand the potential and limitations of SSL in addressing the scanner effect.

1.3 Contribution

Comprehensive Literature Review: In An in-depth examination of the scanner effect, detailing its implications and significance in medical image analysis. This review also encompasses existing scanner-invariant techniques, their advantages, and limitations.

Cross-Domain SSL Experiments: A pioneering study showcasing the performance benefits of using cross-domain self-supervised techniques over domain-specific (natural or medical) techniques, especially when applied to supervised tasks (Dadsetan et al., 2022).

Evaluating Scanner Effect Reduction using different transfer learning: Empirical evidence demonstrating the ability of Cross-Domain SSL techniques to significantly reduce batch effects, establishing its potential as a robust alternative to traditional harmonization methods.

Scanner-Aware SSL Enhancement: A novel methodology that incorporates scanner-specific information during SSL, further enhancing its potential to address the scanner effect.

Comparative Analysis with Unsupervised Harmonization: A detailed comparison between SSL and unsupervised harmonization techniques, highlighting the relative strengths and weaknesses of each approach.

2.0 Background

2.1 Literature Review on Scanner Effect

2.1.1 Definition and understanding of the scanner effect

The "scanner effect", also referred to as the "batch effect" or "site effect" in some literature, pertains to the variability and discrepancies introduced in medical images due to differences in scanning equipment, protocols, and settings (Stonnington et al., 2008; Svanera et al., 2024). This variability arises from multiple factors, including differing manufacturers, models of scanners, imaging parameters, and even software versions used for reconstruction. Besides equipment-related variabilities, patient positioning, and physiological conditions (like blood flow or breathing patterns) during scanning can further exacerbate this effect.

The scanner effect has significant implications for medical image analysis. From a research standpoint, when images from multiple sites or scanners are combined for a study, the inherent discrepancies due to the scanner effect can confound the results. It may overshadow genuine biological or pathological differences, leading to biased conclusions and inferences. Additionally, machine learning models, particularly deep learning models trained on data from a specific scanner, might exhibit degraded performance when applied to data from a different scanner or site. This compromises the generalizability and transferability of algorithms and models, which is crucial for real-world clinical applications.

Consistency in medical images is pivotal. It ensures that the images, regardless of their source or acquisition protocol, reflect accurate and comparable pathological or physiological information. The scanner effect significantly undermines this consistency. For instance, the same tissue might appear differently on images acquired from two different MRI machines due to differences in magnetic field strengths or imaging sequences (Islam et al., 2023; Safari et al., 2024). Such inconsistencies can lead to misinterpretations, especially in longitudinal studies where patient scans are taken at different time points or locations. This lack of consistency necessitates the development of harmonization methods to make multi-site or

multi-scanner data homogenous.

2.1.2 Effects on clinical data interpretation

From a clinical standpoint, the scanner effect has far-reaching ramifications. Different appearances of the same pathology across different scanners can result in diagnostic inconsistencies. A lesion that's clearly visible in a scan from one machine might appear subdued or different in another, leading to potential diagnostic oversight. Such variations also affect quantitative imaging, where precise measurements (like tumor volume or tissue density) are crucial for treatment planning or response evaluation. Inconsistent measurements due to scanner effects can lead to inappropriate clinical decisions. Moreover, for diseases where early detection is crucial, such as in certain cancers, variations introduced by the scanner effect might delay diagnosis and consequently, timely intervention.

2.1.3 Methods to mitigate the effect of scanner-specific variations

MRI scanner invariance aims to correct or mitigate the effects of scanner-specific variations in the acquired MRI data. This is a crucial objective in neuroimaging, ensuring consistent and reproducible findings irrespective of scanner-specific variations. As researchers dig deeper into this challenge, a diverse spectrum of methodologies has emerged to address it.

Starting at the foundational level, pre-processing techniques have long been the bedrock of ensuring consistency (Tudorascu et al., 2016). Methods such as bias field correction (Tustison et al., 2010) rectify intensity inhomogeneities attributable to magnetic field variations. Complementing this, intensity normalization (Bansal et al., 2017; Reinhold et al., 2019) works to equate the intensity scale across MRI outputs from different scanners. Spatial normalization ensures anatomical alignment across datasets by warping images to a universally recognized standard space.

While pre-processing provides a significant rectification level, harmonization techniques have emerged as specialized tools to adjust for scanner effects. Notably, techniques like ComBat (Johnson et al., 2007; Fortin et al., 2018; Beer et al., 2020; Torbati et al., 2021), which originally found applications in genomics, are now adeptly harmonizing MRI data. On

the other hand, MIDAS (Saykin et al., 2010) employs statistical methodologies tailored for multi-site imaging data, and patch-based methods (Tournier et al., 2004; Garyfallidis et al., 2014; Blumberg et al., 2018, 2019) leverage local MRI image patches to correct site-specific discrepancies.

Diving deeper into the realm of data handling, data augmentation techniques such as domain adaptation (Guan and Liu, 2021; Guan et al., 2021) mitigate the domain shift by transferring knowledge from one scanner domain to another. Furthermore, the advent of generative models, especially GANs (Gatys et al., 2015; Isola et al., 2017; Zhu et al., 2017; Sun et al., 2020; Shin et al., 2018; Tomar et al., 2022; Liu et al., 2021), enables the creation of synthetic images that emulate various scanner characteristics, enriching the data pool and enhancing model generalization.

Taking a more model-centric perspective, model-based approaches like multi-task learning (Wang et al., 2020; Ma et al., 2018; Hu et al., 2019) not only predict the clinical outcome but also identify the scanner type, inherently accounting for scanner variations. Another fascinating frontier is feature disentanglement (Zhao et al., 2023; Bayer et al., 2022; Gu et al., 2023; Zuo et al., 2021b) within neural networks, wherein architectures are molded to separate scanner-specific features from the more pertinent anatomical or disease-specific ones.

Even after model deployment, post-processing techniques come into play. Statistical corrections (Singh et al., 2017; Vovk et al., 2007), for instance, recalibrate derived measures post-analysis, while residual analysis (Zhang et al., 2022; Chen et al., 2023) identifies and rectifies scanner-specific patterns lurking within the data.

Broadening the scope to joint analysis techniques (Kurokawa et al., 2021; Tong et al., 2020; Siqueira Pinto et al., 2023), jSBM (Xu et al., 2009) stands out, linking voxelwise and independent component patterns from the same subjects. Meanwhile, multimodal integration (Zhang et al., 2020), which merges data from disparate imaging modalities, has shown promise in accentuating scanner invariance.

In the domain of embedding and representation learning, deep embeddings (Zhu et al., 2018) have surfaced as potential game-changers, learning representations intrinsically resilient to scanner differences. Concurrently, manifold learning (Qiu et al., 2015; Zhu et al.,

2018) techniques like t-SNE offer a different lens, identifying intrinsic data structures and possibly scanner-specific clusters.

Figure 1 illustrates the taxonomy of methods commonly employed to ensure MRI data is scanner invariant.

2.2 Harmonization Techniques: An Overview

2.2.1 Definition and objectives of harmonization

Harmonization in medical imaging refers to the process of adjusting and refining images from diverse sources to ensure that they are consistent, comparable, and standardized. This adjustment transcends varied imaging devices, protocols, and patient populations to produce images that resonate on a universal scale, regardless of where or how they were acquired. Objectives of harmonization include:

Ensuring data comparability. At its core, harmonization seeks to make sure that images from different sources can be analyzed collectively and consistently, eliminating discrepancies caused by differing acquisition parameters.

Boosting research reproducibility. By standardizing images, harmonization aims to bolster the reproducibility of research findings, ensuring that results are consistent across different studies and datasets.

Facilitating cross-modality analysis. Harmonization is pivotal in enabling comparability between different imaging modalities, ensuring that insights drawn from one modality can be coherently mapped to another.

Improving clinical decision-making. For clinicians, harmonized images mean more accurate and reliable data, which may lead to better-informed treatment decisions and patient outcomes.

Optimizing resources. By using harmonization techniques, researchers can efficiently utilize existing datasets without the need for re-acquisition, saving both time and resources.

Upholding data integrity. One of the prime objectives of harmonization is to ensure

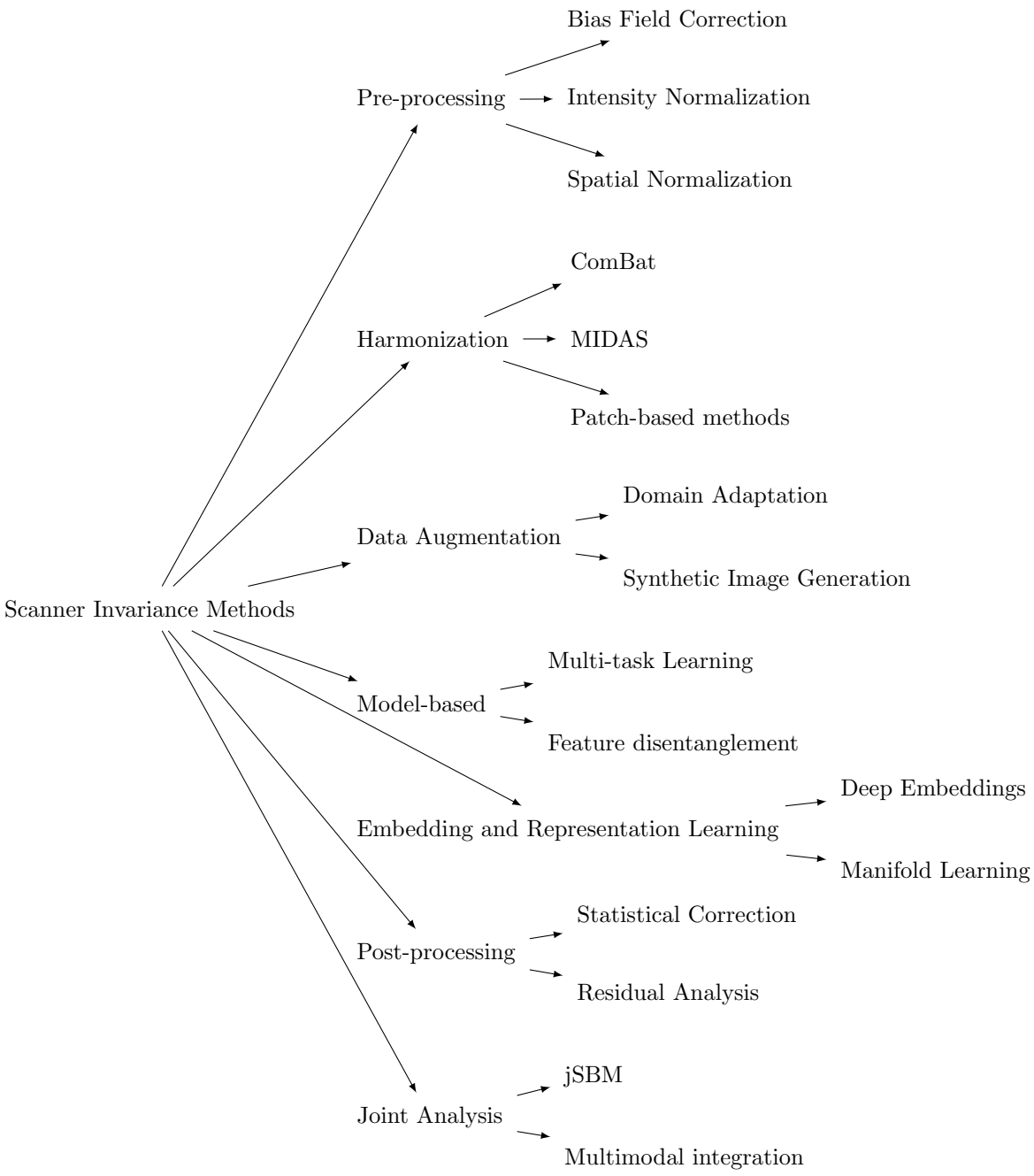


Figure 1: Taxonomy of methods employed to achieve MRI data scanner invariance.

that while making images consistent, the inherent information and details within the images remain undistorted and intact.

As with any evolving technique in diagnostics and research, harmonization in medical imaging brings both numerous advantages and challenges. Understanding them can help researchers and clinicians make informed decisions regarding the adoption and implementation of harmonization protocols. Some challenges that may be introduced by using harmonization techniques are: 1) Risk of Data Over-Manipulation: While harmonization aims to make images comparable, there's a risk of over-manipulating the data, potentially introducing artifacts or losing vital information. 2) Complexity and Technical Challenges: Implementing some harmonization techniques, especially those based on deep learning, requires advanced computational resources and expertise. This might be challenging for smaller institutions or clinics to adopt. 3) Not Always Perfect: While harmonization can considerably reduce variability, it's not always perfect. Some intrinsic scanner-specific characteristics or patient-specific variations might remain unaddressed. 4) Dependence on Reference Standards: Many harmonization techniques require reference standards or phantoms. If these standards are not universally accepted or if they deviate over time, it might lead to inconsistencies. 5) Potential Bias Introduction: Techniques like ComBat, when randomly applied, can potentially introduce biases, especially when the number of images from different sources is imbalanced.

In light of the above, while harmonization in medical imaging undeniably advances the field, offering consistency and improved interpretability, it is not without challenges. These techniques often involve normalization, registration, and other preprocessing steps that seek to bring the data into alignment. However, this pursuit of uniformity can subtly alter the anatomical structures within the images. For example intensity normalization, a vital technique, reduces variations in contrast and brightness across images. But its gentle touch can reshape tissue contrast, potentially affecting the depiction of anatomical features. In another example, geometric distortion correction enhances spatial accuracy by rectifying distortions. Yet, it may inadvertently tweak the shape and size of anatomical structures, introducing variations in measurements and interpretations. Likewise, image registration aligns images spatially but often involves nonlinear transformations that can subtly deform anatomical structures. This quest for alignment can reshape structures, challenging the preservation of anatomical fidelity

2.2.2 Related works

Given the diversity in imaging devices, protocols, and patient cohorts, the presence of inherent variability in acquired images is almost inevitable. To foster a universal interpretation and comprehensive analysis of these images, harmonization techniques emerge as the lynchpin. This chapter presents an exploration of the taxonomy of these techniques, highlighting their significance and potential applications.

Statistical techniques. At the heart of harmonization lie statistical methods, which ensure that images from varied sources resonate on a similar frequency. The method of Histogram Matching stands out, offering an adjustment of the intensity distribution of a source image to echo that of a reference image. Such alignment guarantees that the images possess analogous intensity distributions, paving the way for more accurate comparative analyses (Pizer et al., 1987; Fortin et al., 2016; Shinohara et al., 2014). Another noteworthy technique is ComBat, a method traditionally associated with genomics. By zeroing in on scanner-specific biases, ComBat showcases its efficacy in curtailing unwanted variability, especially those birthed by different scanning sites or devices (Johnson et al., 2007). Z-score Normalization simplifies this endeavor, transforming images to maintain a standard scale and distribution (Nyúl et al., 2000).

Correcting retrospectively. Often, images acquired in the past need retrospective adjustments for contemporary applications. Techniques like Bias Field Correction come to the fore in such scenarios, especially when dealing with MRI images riddled with spatial intensity variations (Ashburner and Friston, 2005). The N4ITK method takes this a notch higher, focusing on refining intensity non-uniformities in MR images, making it especially effective for brain imaging (Sled et al., 1998).

Deep learning. The advent of deep learning has ushered in a revolutionary phase in image harmonization. Convolutional Neural Networks (CNNs), with their multilayered architecture, adeptly extract and harmonize image features, adjusting for variations across large datasets. The ingenuity of Generative Adversarial Networks (GANs) (Zhu et al., 2017) in synthesizing harmonized images, thanks to its dual neural networks, further cement deep learning's pivotal role in this domain (Dar et al., 2019; Kieselmann et al., 2021; Zhong et al.,

2020). Autoencoders, particularly their variant - variational autoencoders, offer another layer of sophistication, encoding and decoding images to ensure harmonization (Zuo et al., 2021b; Fatania et al., 2022).

Domain adaptation and multivariate techniques. Harmonizing images across different domains necessitates specialized techniques. Joint Distribution Adaptation (JDA) (Zuo et al., 2021a) and Transfer Component Analysis (TCA) (Guan et al., 2021) shine in this respect, minimizing distribution divergences and finding domain-invariant spaces, respectively. Beyond these, multivariate techniques like Canonical Correlation Analysis (CCA) (Bashyam et al., 2020) and Partial Least Squares (PLS) Lebedev et al. (2013) work diligently to align datasets into a shared, harmonized space.

The prospective angle. While most techniques address discrepancies post-acquisition, prospective harmonization adopts a proactive approach. Implementing Standardized Imaging Protocols ensures that every image, regardless of its origin, conforms to a predefined standard. Phantom-based Calibration (Timmermans et al., 2019; Karayumak et al., 2019; Keenan et al., 2018), which involves calibrating scanners using physical models, further solidifies the quest for uniformity across different machines.

Distinguishing harmonization as supervised or unsupervised technique. Distinguishing between the supervision requirements of these techniques, methods like Histogram Matching, Affine Registration, and Z-score Normalization operate in an unsupervised manner. Deep learning techniques, including certain CNN and GAN models, often require supervised training, leveraging labeled datasets. However, autoencoders and some GAN variants (like CycleGAN) can be trained in an unsupervised manner.

2.3 Self-Supervised Learning: A New Approach

Self-supervised learning (SSL) represents a growing domain in the landscape of machine learning, especially with its applications in the realm of deep learning. While supervised learning has been the dominant paradigm, requiring labeled datasets for training, SSL introduces a paradigm shift by exploiting the inherent structure of the data itself. In this

methodology, the system is trained to predict parts of the data from other parts, thereby generating its own supervisory signal. This enables a vast exploitation of the available unlabeled data, setting SSL apart and providing it with a potent edge in diverse applications, notably in medical imaging where labeled data is often scarce and expensive to acquire.

2.3.1 Benefits over traditional harmonization

Traditional harmonization methods in medical imaging, whether statistical or deep learning-based, have primarily depended on curated and labeled datasets. However, the generation of such datasets is labor-intensive, often requiring expert annotations which may be susceptible to subjective errors. SSL offers a compelling alternative by leveraging the vast amounts of available unlabeled medical images.

One notable advantage is the potential for improved model generalization. By training on a broader spectrum of data, models can learn more representative features that ensure robust performance across varied datasets. Additionally, SSL can be more scalable. As medical imaging continues to generate massive volumes of data, the ability to utilize this data without the need for labeling translates to more agile and timely model training and deployment.

Moreover, SSL dovetails with transfer learning, wherein a pre-trained model on a large dataset can be fine-tuned for specific tasks using smaller labeled datasets. This synergy ensures that even in scenarios where labeled data is available, but in limited quantities, SSL can bridge the gap, enhancing performance and reducing the need for extensive labeled data.

2.3.2 Self-supervised learning and scanner effect

A recurrent challenge in medical imaging is the batch or scanner effect, where discrepancies arise due to variations in scanner models, protocols, or even site-specific idiosyncrasies. Traditional methods often required explicit modeling of these effects, leading to intricate pipelines that might not fully encapsulate the nuanced batch variations.

SSL offers a nuanced understanding of these effects. As SSL models are trained on the inherent structure of the data, they can potentially detect and account for scanner-specific

patterns, ensuring consistent feature extraction across scanners. Studies have shown (Jiang et al., 2021; Chang et al., 2022; Dhinagar et al., 2023; Kan et al., 2022) that representations learned through SSL are often more invariant to such effects, ensuring that the subsequent tasks, whether classification or regression, are less affected and more focused on anatomical information.

Furthermore, the SSL paradigm facilitates the creation of domain-adaptive models (Kan et al., 2022). When deployed in multi-center studies or large-scale clinical deployments where data from diverse scanners is aggregated, the robustness imparted by SSL can be invaluable, ensuring consistent and reliable outcomes irrespective of the source of the data.

2.3.3 Importance of cross-domain learning: Natural vs. Medical Images

Artificial intelligence requires efficient deep learning techniques and large amounts of training data to develop reliable and robust systems. As a result of the complexity of annotation tasks and the high level of expertise needed for manual interpretation, the construction of labeled datasets is often time-consuming and expensive, such as in the medical imaging domain. Transfer learning from natural images is becoming increasingly popular in medical imaging to overcome the lack of annotations. (Liu et al., 2020b; McKinney et al., 2020; Menegola et al., 2017; Xie et al., 2019). Although numerous experimental studies indicate the effectiveness of fine-tuning from either supervised or self-supervised ImageNet models((Alzubaidi et al., 2020; Graziani et al., 2019; Heker and Greenspan, 2020; Zhou et al., 2021; Hosseinzadeh Taher et al., 2021; Azizi et al., 2021)), it does not always improve the performance due to domain mismatch problem (Raghu et al., 2019).

Furthermore, SSL has demonstrated great success in many downstream computer vision applications, where labeling is time-consuming and expensive (Doersch et al., 2015; Gidaris et al., 2018; Noroozi and Favaro, 2016; Zhang et al., 2016; Ye et al., 2019; Bachman et al., 2019; Tian et al., 2020a; Henaff, 2020; Oord et al., 2018). Due to the enormous volume of medical images generated by clinical and research settings, SSL learning approaches are particularly well suited to medical research and healthcare. However, SSL approaches have received limited attention in the medical image domain despite this demand. A few studies

have examined the role of SSL for medical image analysis for only a limited number of applications, including classification (Liu et al., 2019; Sowrirajan et al., 2021; He et al., 2020b; Azizi et al., 2022; Zhu et al., 2020; Liu et al., 2020a) and segmentation (Ronneberger et al., 2015; Bai et al., 2019; Chaitanya et al., 2020; Spitzer et al., 2018).

Our research focuses on developing a self-supervised deep learning algorithm for predicting the progression of Alzheimer's disease (AD) through the use of high-dimensional magnetic resonance imaging (MRI). Patients with AD show clinical symptoms years after onset of the disease, due to the slow degeneration of brain cells. Consequently, preventing irreversible and fatal brain damage requires an accurate diagnosis and treatment of AD in its early stages. With accurate prediction of AD progression, clinicians can start treatment earlier and provide more personalized treatment.

In order to predict AD progression, many existing methods have categorized patients into coarse categories, such as Mild Cognitive Impairment (MCI) or dementia. These methods have also been used to predict progression from one category to another (e.g. MCI to dementia) (Risacher et al., 2009; Venugopalan et al., 2021; Oh et al., 2019). Clinical trials, however, require finer-grained measurement scales since trial populations tend to be narrowly defined (e.g. only MCI patients). Alternatively, continuous numerical values can be used to predict the outcome of cognitive and functional tests. By framing the prediction task as a regression rather than a classification, prognostic models offer more granular estimates of disease progression. Recently, a few deep learning-based approaches, including the recurrent neural network (RNN) and convolutional neural networks (CNN) have been proposed for predicting disease progression of AD patients based on MRIs. (Nguyen et al., 2020) adapted MinimalRNN to integrate longitudinal clinical information and cross-sectional tabular imaging features for regressing endpoints. (El-Sappagh et al., 2020) utilized an ensemble model based on stacked CNN and a bidirectional long short-term memory (BiLSTM) to predict the endpoints on the fusion of time series clinical features and derived imaging features. Recently, several methods have started to employ CNN-based models to extract features from raw medical imaging. (Tian et al., 2022) applied CNN with multi-task interaction layers composed of feature decoupling modules and feature interaction module to predict the disease progression.

Despite demand, little progress has been made because of the difficult design requirements, lack of large-scale, homogeneous datasets that contain early-stage AD patients, and noisy endpoints that are potentially hard to predict. Much of the prior work has focused on using image-derived features to overcome the complexity and high variability in raw MRIs and small datasets. Most current prognosis models are trained on a single dataset (i.e. cohort), which limits their generalizability to other cohorts. They also use a limited number of annotated images, which can lead to problems such as domain shift and heterogeneity.

We establish a cross-domain self-supervised transfer learning approach that learns transferable and generalizable representations for medical images. Our approach leverages SSL on both unlabeled large-scale natural images and an in-domain medical image dataset comprised from 5 different studies. These representations can be further fine-tuned for downstream tasks such as disease progression prediction, using limited labeled data from the clinical setting. We evaluate the performance of different supervised and self-supervised models pretrained on either natural images or medical images, or both. Our extensive experiments reveal that (1) Self-supervised pretraining on natural images followed by self-supervised learning on unlabeled medical images outperforms alternative transfer learning methods, indicating the potential of SSL in reducing the reliance on data annotation compared to supervised approaches (2) Self-supervised models pretrained on medical images outperform those pretrained on natural images, denoting that SSL on medical images yields discriminative feature representations for regression task.

2.3.3.1 Related Works

The recent advancements and achievements in self-supervised learning techniques, such as contrastive learning (Wu et al., 2018; He et al., 2020a; Chen et al., 2020c,b,a; Grill et al., 2020; Misra and Maaten, 2020), mutual information reduction (Tian et al., 2020b), clustering (Caron et al., 2020; Li et al., 2020), and redundancy-reduction methods (Zbontar et al., 2021; Bardes et al., 2021) in the field of computer vision highlight their effectiveness in enhancing the performance of Artificial Intelligence (AI) systems. These techniques involve training models on various pretext tasks to enable the network to acquire high-quality

representations without relying on label information. One example is SimCLR (Chen et al., 2020c) which aims to maximize agreement between representations of different augmentations of the same image by using a contrastive loss in the latent space. Another method, Barlow Twins (Zbontar et al., 2021), measures the cross-correlation matrix between the embedding of two identical networks, with the goal of making this cross-correlation close to the identity matrix. Meanwhile, SwAV (Caron et al., 2020) simultaneously clusters the images while enforcing consistency between cluster assignments produced for differently augmented views of the same image, rather than comparing features directly as in contrastive learning.

Subsequently, SSL has been employed for medical imaging applications including classification and segmentation to learn visual representations of medical images by incorporating unlabeled medical images. While some approaches have designed domain-specific pretext tasks (Bai et al., 2019; Spitzer et al., 2018; Zhuang et al., 2019; Zhu et al., 2020), others have adjusted well-known self-supervised learning methods to medical data (He et al., 2020b; Li et al., 2021; Zhou et al., 2020; Sowrirajan et al., 2021). Very recently (Azizi et al., 2022) has applied SimCLR on a combination of unlabeled ImageNet dataset and task-specific medical images for medical image classification; their experiments and improved performance suggest that pretraining on ImageNet is complementary to pretraining on unlabeled medical images.

Although aforementioned approaches demonstrate improvement of the performance on challenging medical datasets, all of them are limited to classification and segmentation tasks and their benefits and potential effects for the prognosis prediction tasks, as regression tasks, have not been studied. Formulating progress prediction as a regression rather than a traditional classification problem leads to a more fine-grained measurement scale which is crucial for real-world applications. Therefore, the development of self-supervised networks is in great demand for efficient data utilization in medical imaging for disease prognosis. To the best of our knowledge, this is the first study of developing a self-supervised deep convolution neural network on medical data images from various cross-domain datasets to predict a granular understanding of disease progression.

3.0 Materials and Methods

3.1 Task and Data

In Alzheimer's Disease clinical trials, one of the most important cognitive tests used to assess current patient function and the likelihood of AD progression is Clinical Dementia Rating Scale Sum of Boxes (CDR-SB). CDR-SB is a score provided by clinicians based on clinical evaluations and its ranges from 0 to 18, with higher scores indicating greater severity of symptoms. CDR-SB score is then used to assign Alzheimer's status of a patient.

Our goal is to use a regression approach to predict the future status of patients with AD based on their initial visit. Specifically, our model takes as input 2D slice stacks of an MRI volume that are collected at the first visit and predicts the CDR-SB value at month 12 (i.e. after one year). By accurately predicting the progression of AD in patients, clinicians can initiate treatment at an earlier stage and tailor the most suitable and effective treatment for each individual patient.

All individuals included in our analysis are around $5k$ from five studies, including ADNI (Petersen et al., 2010), AIBL (Ellis et al., 2009), HABS (Dagley et al., 2017), OASIS-3 (LaMontagne et al., 2019), and WRAP (Langhough Koscik et al., 2021).

The following steps are used to standardize MR volumes. The first step is to infer a brain mask using SynthSeg (Billot et al., 2021), a deep learning segmentation package. We resample the volumes and segmentations isotropically to 1 mm voxel size, standardize the orientation to canonical (RAS+), rescale the intensity to 0.1, and normalize the Z-score during training. As a final step, volumes are cropped or padded to (224,224).

We prepare 5-slice dataset medical images for training self-supervised models. 5-slice dataset means that from each 3D MRI volumes, five slices are extracted to create the 5-slice. This stack consists of the middle slice of the brain sub-volumes and four adjacent slices, with intervals of 5 (two to the right, two to the left). SSL training sets include subjects from all datasets except OASIS-3, which are reserved as out-of-study test sets (refer to Table 1). The 5-slice dataset contains a total of $122,245$ unlabeled images. Approximately 90% of the

Table 1: Summary of datasets. The ✓ indicates whether a study is utilized for a split. OASIS-3 is designated as out-study test sets, meaning they have not been utilized for either SSL pretraining or fine-tuning. The in-study test set includes patients from ADNI; but there is no overlap between the splits. We are unable to find any labels for the first 3 rows of datasets.

Study	Number of patients	SSL	Fine-tuning	in-study test	out-study test
HABS	289	✓	-	-	-
AIBL	1112	✓	-	-	-
WRAP	578	✓	-	-	-
ADNI	2332	✓	✓	✓	-
OASIS-3	46	-	-	-	✓

development set is used for training, while the remaining 10% is reserved for validation. We choose the best model based on the minimum self-supervised validation loss and transfer its backbone weights to the supervised model.

To create labeled datasets for fine-tuning, we selected participants from ADNI. The fine-tuning dataset comprises about 1000 images, none of which are utilized for self-supervised learning training. Approximately 30% of the fine-tuning dataset is set aside as an in-study test set, while the rest of the data is divided into training and validation sets (see Table 1).

To study the effects of different scanners, we extract three specific details from each DICOM file: the manufacturer, model, and institution associated with the scan. Notably, the institution attribute is anonymized for most of the studies incorporated in our research. Given this limitation, and to consistently use all 4 studies in our self-supervised learning approach and the 2 datasets for supervised tasks (1 for training and validation, 2 for testing), our analysis primarily revolves around the manufacturer and model details.

Figure 3 showcases the distribution of MRI scans across four manufacturers: GE, Philips, Siemens, and Toshiba. A significant portion of our self-supervised learning dataset originates

from Siemens, accounting for approximately 43K MRIs. This is followed by GE with around 24K MRIs, Philips with about 17K MRIs, and Toshiba contributing fewer than 1K MRIs. The breakdown of different scanner models under each manufacturer is depicted in Figure 2.

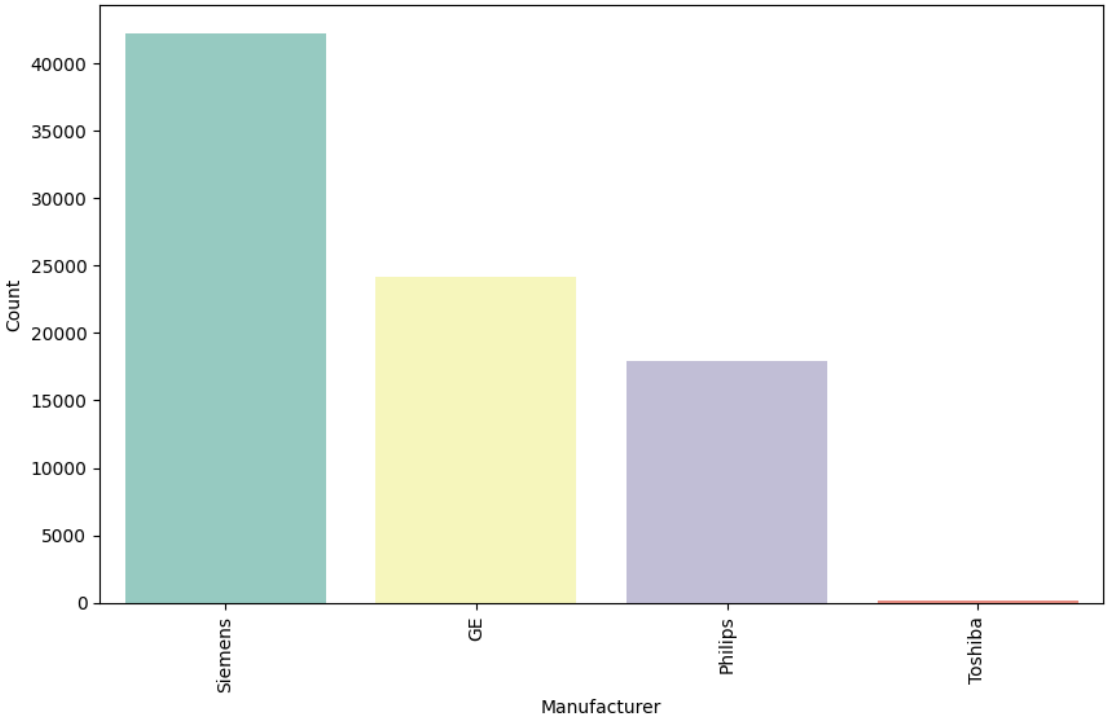


Figure 2: Bar plot showing the distribution of 4 different manufacturers across all self-supervised learning studies including Siemens, GE, Philips and Toshiba.

For the fourth and fifth contributions of this thesis, we increase the number of data points for the fine-tuning part of our model. This section presents an analysis of the new training and validation datasets, highlighting the distribution and characteristics of the data. The training data has been significantly expanded to improve the model's performance. Figure 4 illustrates the distribution of the CDR-SB variable for baseline and 12-month follow-up.

Histogram of CDR-SB at baseline. Based on Figure 4 the distribution is heavily skewed towards lower values, with a peak frequency at CDR-SB of 0 and a gradual decline as CDR-SB increases but at 12-month visit, this distribution is more spread out than the baseline, with a noticeable peak around CDR-SB equal to 2 and a long tail extending to

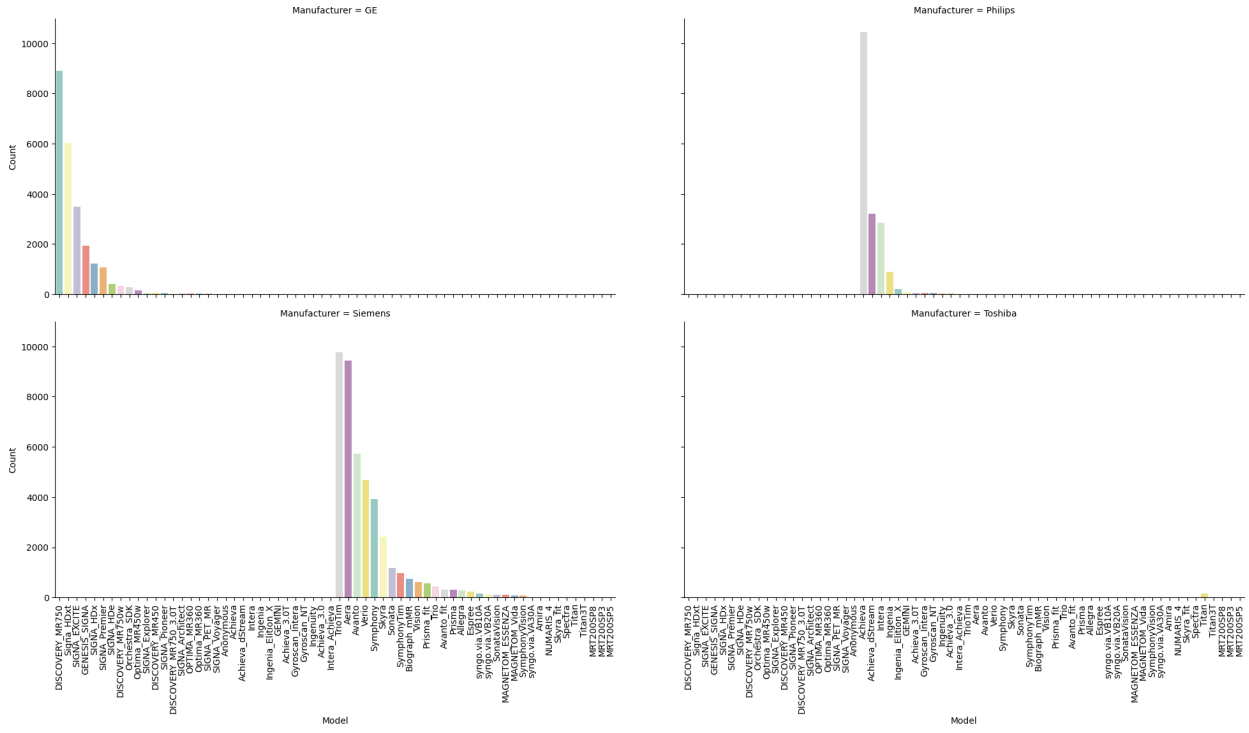


Figure 3: Distribution of scanner models within each manufacturer.

higher values.

The validation data, used to provide an unbiased evaluation of the model during fine-tuning, shows similar characteristics to the training data, ensuring consistency and reliability in model evaluation.

Figure 5 shows the validation data histograms at both baseline and 12-month follow-up. The distribution of CDR-SB at baseline is similar to the training data, with a peak at lower CDR-SB values. Also, the distribution of CDR-SB at follow-up mirrors the training data, with a peak around CDR-SB of 2 and a long tail.

The expanded training and validation datasets provide a robust foundation for fine-tuning the model. The consistency in distributions between the training and validation sets ensures that the model can be evaluated accurately and adjusted effectively, leading to improved performance and generalization.

By studying Figure 6 and Figure 7, the training and validation datasets exhibit similar

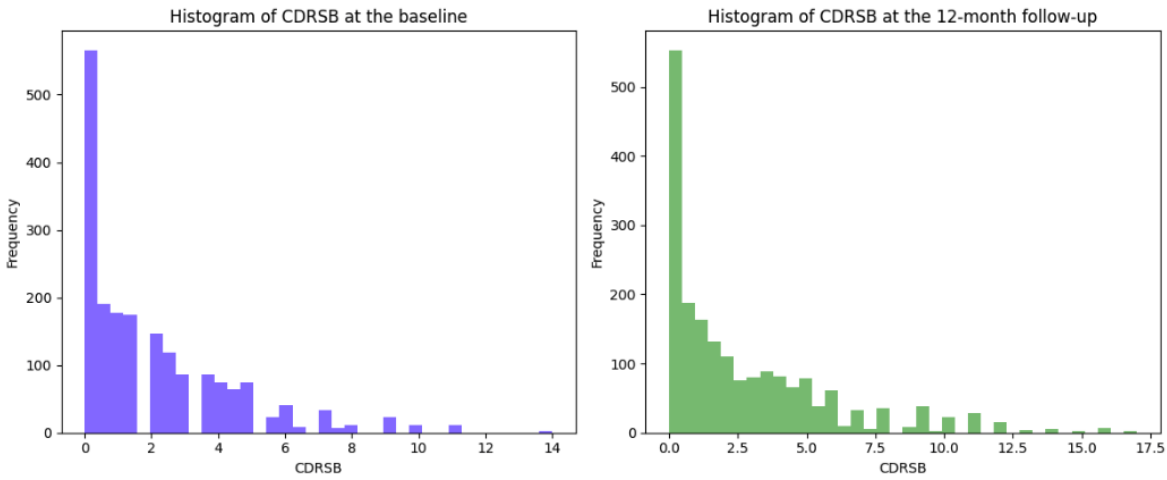


Figure 4: This figure illustrates the distribution of the CDR-SB variable for baseline and 12-month follow-up for the training dataset.

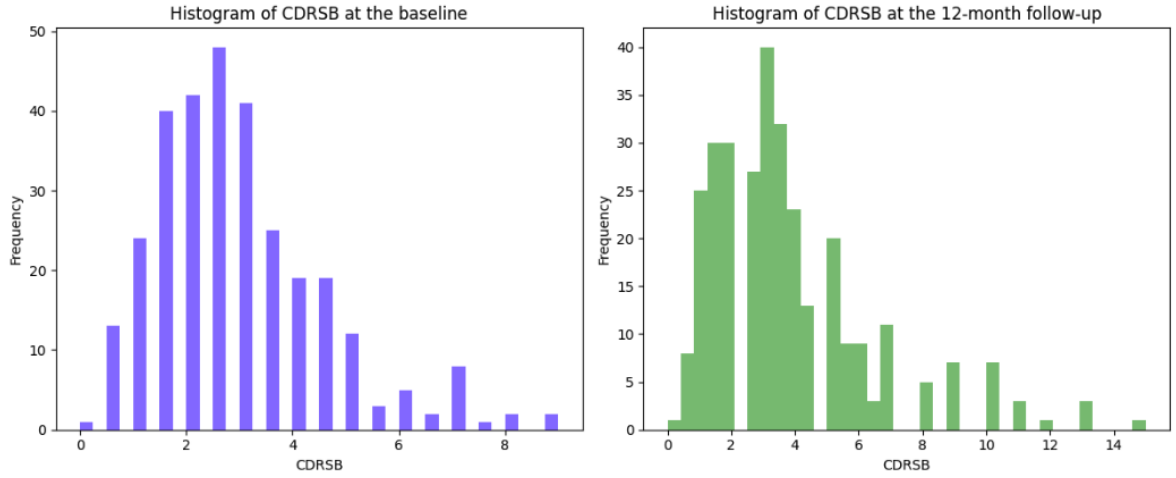


Figure 5: This figure illustrates the distribution of the CDR-SB variable for baseline and 12-month follow-up for the validation dataset.

patterns across various categorical distributions, indicating consistency between the two sets. In both datasets, the model distribution shows GE SIGNA as the most frequent model, followed by a range of other models with decreasing frequencies. This diversity in

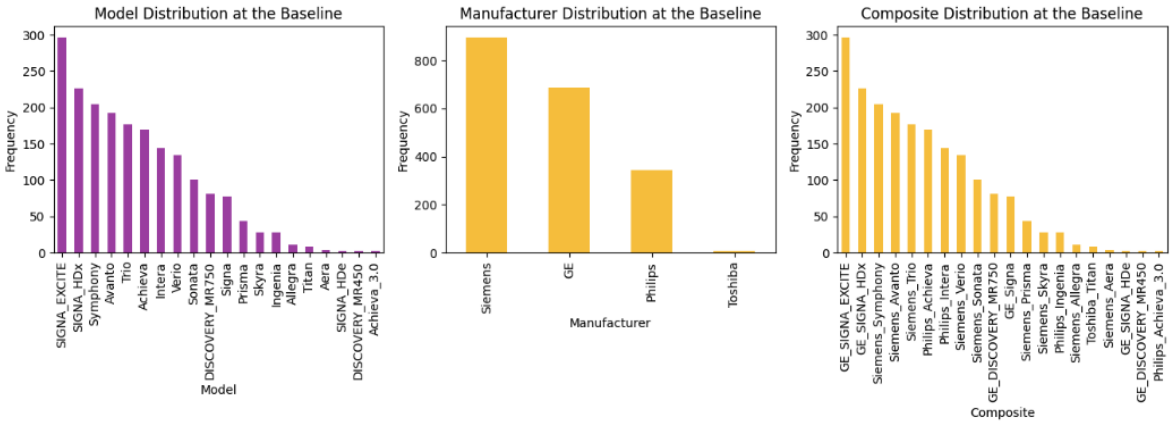


Figure 6: Distribution of models, manufacturers, and their composition at the baseline for the training dataset.

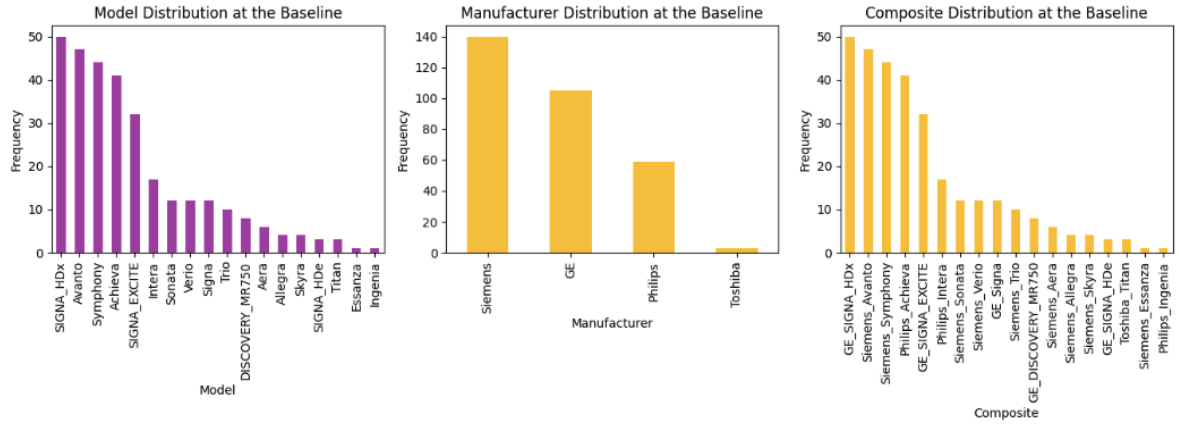


Figure 7: Distribution of models, manufacturers, and their composition at the baseline for the validation dataset.

models suggests a broad representation that could contribute to a robust training process. The manufacturer distribution in both sets is dominated by Siemens, GE, and Philips, with Siemens having the highest frequency. This consistency across training and validation data ensures reliable model evaluation on different equipment manufacturers. The composite distribution also mirrors this pattern, with SIGNA being the most common in both sets.

Also by running Chi Square experiments, we find that the p-value of the difference between manufacturer distribution of training and validation dataset is 0.53. This cannot reject the null hypothesis suggesting no difference in manufacturer distribution between the training and validation dataset.

Table 2 presents the results of an ANOVA test comparing the mean of CDR-SB between different manufacturers for both training and validation datasets.

Table 2: ANOVA test results comparing mean of CDR-SB between different manufacturers.

Metric	Train	Validation
df	3.0	3.0
F	7.15	1.17
P-value	$> 0.0001\,(0.00008)$	0.318

Table 3: ANOVA test results comparing the mean of CDR-SB of training and validation for each dataset between different manufacturers.

P-value	Train	Validation
ADNI	0.0523 (df = 2)	0.942 (df = 2)

For the training dataset, the F-value of 7.15 indicates that there is a significant difference in the means of CDR-SB between the different manufacturers in the training dataset. The very low P-value ($< 0.0001, (0.00008)$) confirms that this difference is statistically significant.

For the validation dataset, the F-value of 1.17 suggests much weaker evidence of a difference in the means of CDR-SB between the manufacturers. The P-value of 0.318 indicates that this difference is not statistically significant.

In summary, the ANOVA test results suggest that there is a statistically significant difference in the means of CDR-SB between different manufacturers in the training dataset but not in the validation dataset. This may imply that the observed differences in the training

dataset do not generalize well to the validation dataset, suggesting potential overfitting or differences in data distribution between the training and validation sets. If we limit the training and validation dataset to looking at the ADNI dataset separately the p-value for the training dataset is 0.052 and for the validation dataset is 0.94 suggesting that there is no statistically significant difference in the means of CDR-SB in the ADNI dataset.

3.2 Experimental Study: Cross-Domain Self-Supervised Learning

3.2.1 Self-supervised learning platform for progression prediction task

We evaluate the performance of three SSL pretraining approaches in predicting disease progression. The first approach is to explore the pretrained models on unlabeled natural images to see if they can be transferred to medical images. The second approach is to use pretrained models on unlabeled in-domain medical images to assess their performance on disease progression. The third approach is to apply cross-domain SSL (referred to as CDSSL) to leverage unlabeled data from multiple domains, including natural images and medical images. To establish a reference point, the target model is trained using random initialization, serving as a baseline for comparison.

Exploring pretrained models on unlabeled natural images. SSL models are trained on large datasets of natural images, such as ImageNet, and have been shown to outperform supervised ImageNet models on several computer vision tasks. In this experiment, we hypothesize that these models could be transferred to medical images and used to predict disease progression. We initialize the backbone encoder with weights from SSL models trained on ImageNet to exploit these benefits. In our study, we specifically concentrate on three prominent SSL methods: SimCLR, a contrastive approach; BarLow Twins (BLT), a redundancy reduction approach; and SwAV, a clustering-based contrastive learning approach. These methods have demonstrated remarkable performance on benchmarks designed for natural images. Although there are other SSL strategies available, their performance on ImageNet is comparable to the ones we have selected.

Exploring pretrained models on unlabeled in-domain medical images. As shown in Figure 8(a), we use SimCLR, Barlow Twins, and SwAV to learn distinctive representations of unlabeled medical images. These methods have all been shown to be effective in the classification and segmentation of medical images. We hypothesize that these models would be able to learn the features that are specific to medical images and be more effective at predicting disease progression than the models trained on unlabeled natural images.

Exploring CDSSL pretrained models on both domains. Representations learned from natural images may not be optimal for the medical imaging domain because of the large distribution shift between natural and medical images. Medical images are typically monochromatic and have similar anatomical structures, while natural images are typically colorful and have a wider variety of objects and scenes. We hypothesize that this discrepancy could be minimized by further pretraining on medical data. As shown in Figure 8(b-c), we use SimCLR, Barlow Twins, and SwAV to learn distinctive representations of unlabeled medical images on top of pretraining on ImageNet.

Fine Tuning. The progression prediction task utilizes a ResNet50 backbone (He et al., 2016) followed by a linear layer, with the backbone being initialized randomly or with pretrained models. The model loss is calculated using the mean square error (MSE) criterion (see Figure 8(d)).

3.3 Using SSL to Address Scanner Effect

3.3.1 Evaluating the efficacy of CDSSL in reducing scanner effect

In order to find whether or not transfer learning from cross-domain SSL helps to reduce the batch effect or not, we design a model that fuse the scanner manufacturer and scanner model to the supervised framework embedding to find whether the prediction of our outcome is scanner invariant or not. In order to do that, we perform a log-likelihood ratio on the prediction of two models. As illustrated in Figure 9 Model 1 is a null model that only receives the MRIs whereas Model 2 as a hypothesis receives both MRIs and scanner information with

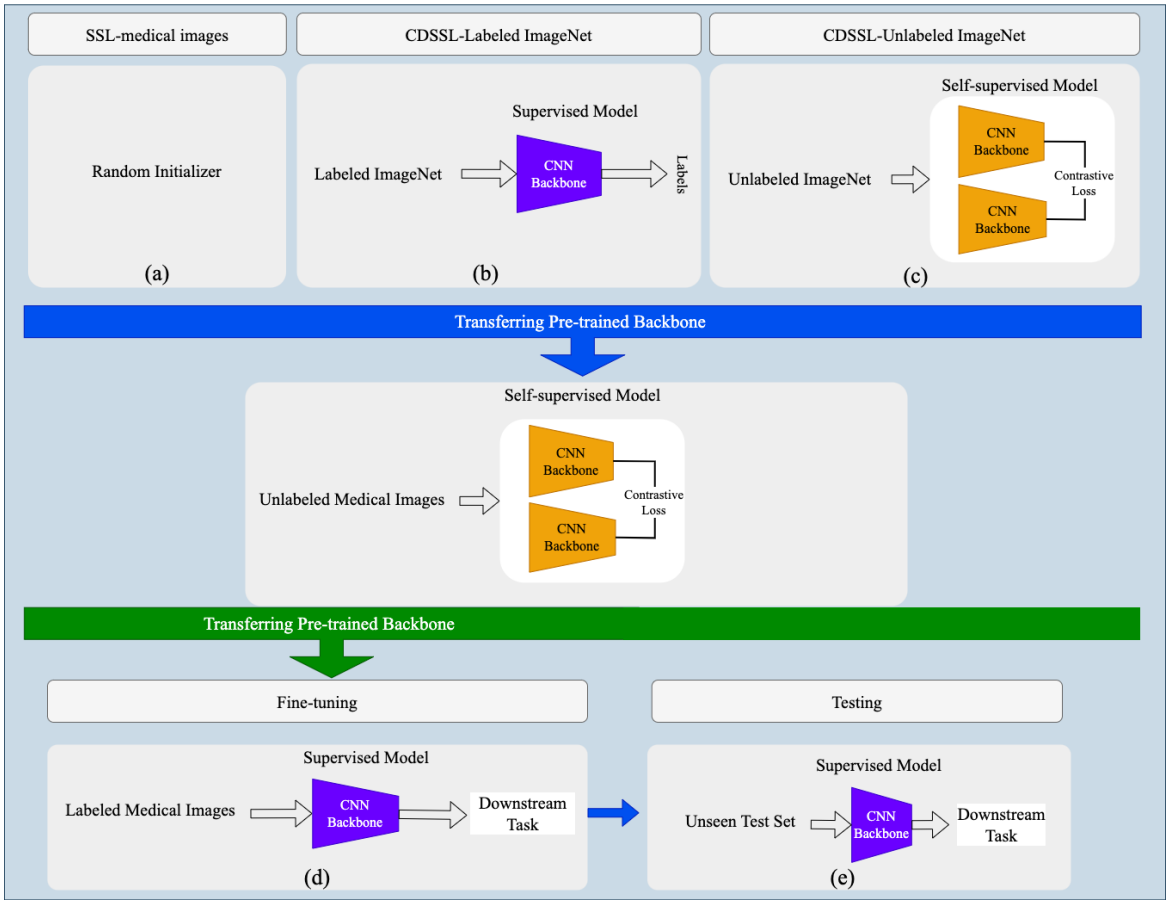


Figure 8: Different approaches for self-supervised pretraining on in-domain medical imaging, including (a) random initialization, (b) supervised ImageNet initialization, and (c) self-supervised ImageNet initialization. (d) Performing fine-tuning by transferring the backbone from one of the scenarios a-c. (e) utilization of the trained model on unseen test sets.

the process of output-fusion. Model 1 represents the simpler model with the assumption that there is no significant effect of MRI scanner variations. Model 2 represents the more complex model that explicitly accounts for MRI scanner effects. We perform a statistical analysis using $p<0.05$ as a statistical significance difference. Noted, in Figure 9-b the process of output-fusion is as follows, first scanner manufacturer and scanner models are processed into one hot vector and then concatenate with the last fully connected layer of ResNet50 (the backbone of our supervised model). Then the combined vector followed by the linear

layer outputs the final prediction.

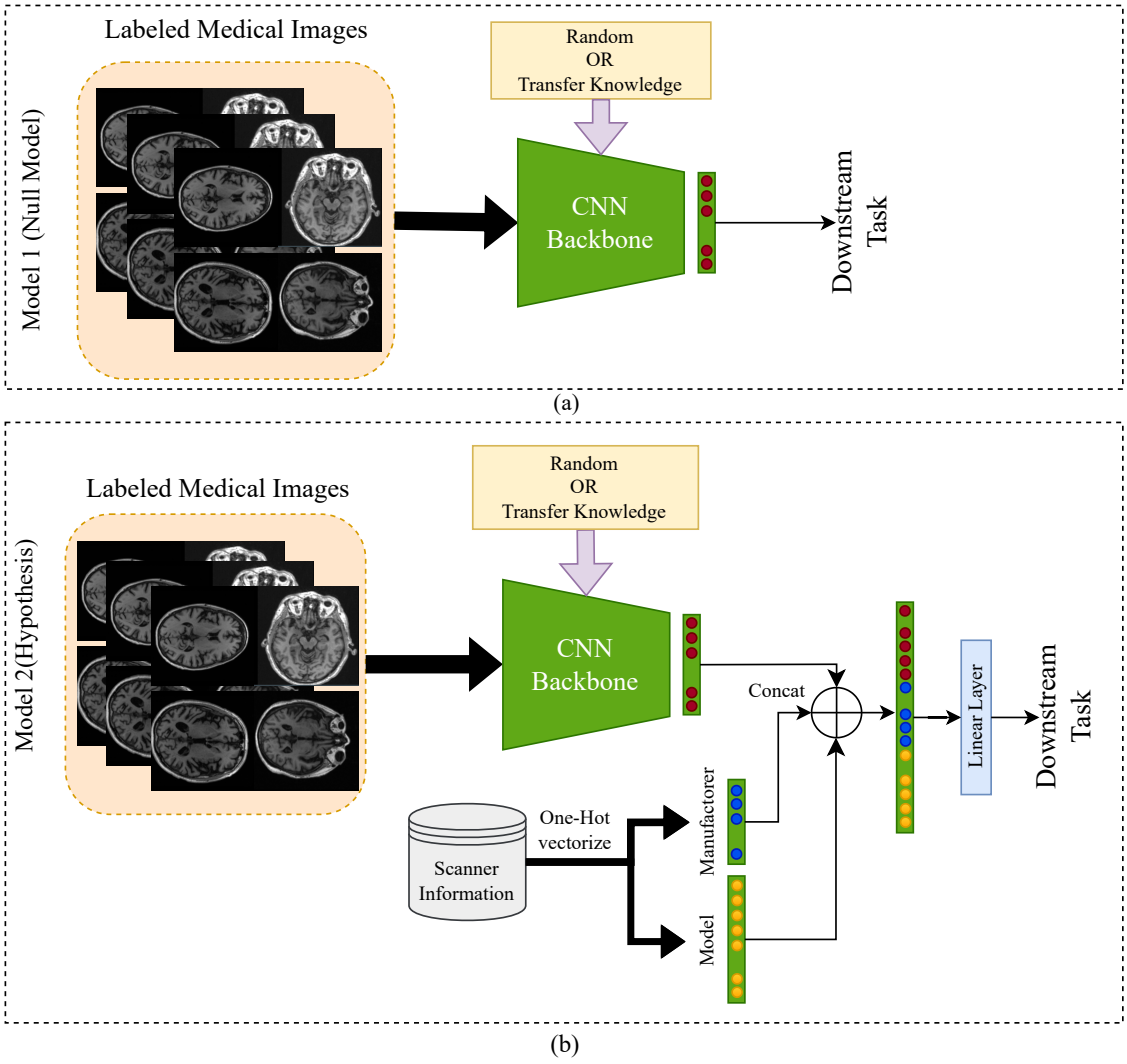


Figure 9: This figure illustrates two models including (a) Model 1 (null model) and (b) Model 2 (hypothesis) with the difference of including scanner information (manufacturer and model) in Model 2.

3.3.2 Enhancing CDSSL with scanner information incorporation

In this section, we are proposing some techniques to introduce scanner-aware self-supervised learning. In the following paragraph, we will describe the details of these proposed methods:

Augmentation strategy. In this strategy, we design data augmentation techniques that mimic potential variations introduced by different MRI scanners. These augmentations include variations in intensity, noise, resolution, and other factors. This approach is partially applied in the previous section on CDSSL, where we used random Gaussian blurring and random rotation. In this section, we extend this strategy by incorporating additional noise and intensity variations and evaluate their impact on the final outcome. We refer to this strategy as "Augmentation CDSSL" throughout this thesis. Key Components of Augmentation CDSSL are:

Mimicking scanner variations

- Intensity Variations: Adjusting the brightness and contrast of the images.
- Noise Variations: Introducing random noise to simulate scanner-specific artifacts.
- Resolution Variations: Modifying the resolution to reflect differences in scanner quality.

Randomized augmentation application

- During each training epoch, images are randomly augmented to introduce variability.
- For example, an image during epoch j might be augmented with random contrast, while in epoch $j+1$, the same image might be used with no augmentation.

Integration with CDSSL

- These augmentations are applied before the default SSL augmentations to create diverse views of the data.
- This strategy enhances the model's ability to generalize across different scanner environments.

Because we have prior knowledge about the number of manufacturers and models, this could guide the choice of clusters and then incorporate that cluster information in the loss function depicted in Equation 1. Then To include the cluster information we define two terms, *Within-cluster Similarity* which increases the similarity of embeddings that belong to the same cluster, and *Across-cluster Dissimilarity* which penalizes high similarity between embeddings from different clusters. Therefore the modified loss becomes:

$$L^{'} = L + \lambda_1 \sum_{m,n\epsilon \text{ same cluster}} sim(z_m, z_n) - \lambda_2 \sum_{m,n\epsilon \text{ different clusters}} sim(z_m, z_n) \qquad (1)$$

Where L is the SimCLR's original Loss and λ_1 and λ_2 are hyperparameters that determine the weight of the within-cluster and across-cluster terms, respectively. During training the model using this modified loss. The standard term pushes the embeddings of the anchor and its positive pair to be similar, while the additional terms make the embeddings of images from the same cluster closer and those from different clusters more distinct. We choose λ_1 as 0.4 and λ_2 as 0.6.

Scanner label prediction as an additional auxiliary task strategy. During training, along with minimizing the contrastive loss for augmented views of the same sample, we introduced an auxiliary task where the network also predicts the scanner (or protocol) that produced the MRI image. This will make the network aware of the scanner variations and could potentially increase its robustness to these variations. We will use "Auxiliary CDSSL" to address this strategy.

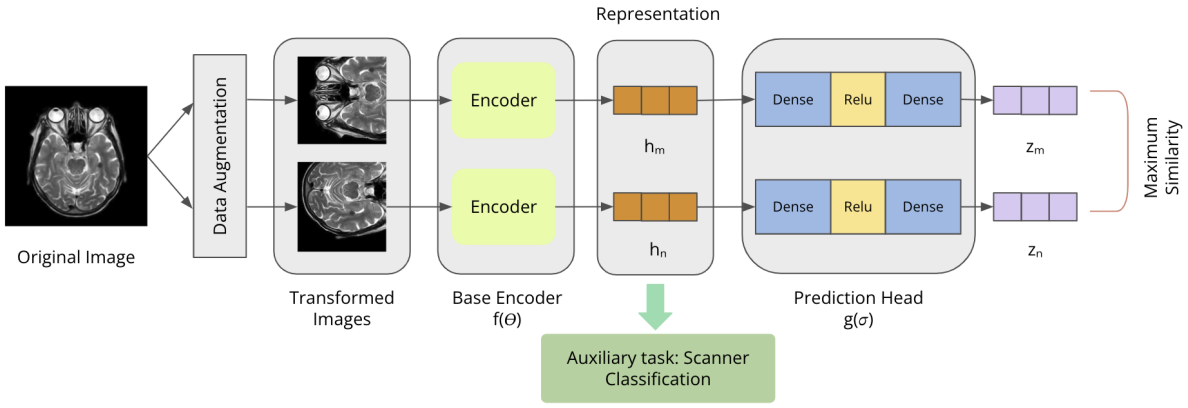


Figure 10: Diagram of the Auxiliary CDSSL method: The original MRI image undergoes data augmentation to create transformed images, which are then processed by the base encoder to generate representations. These representations are fed into the prediction head to maximize similarity. An auxiliary task of scanner classification is incorporated to mitigate scanner variability.

3.3.3 Comparative analysis with unsupervised harmonization technique

To determine if using self-supervised learning (SSL) is superior to harmonization techniques, we performed an analysis where we harmonized all the scanners into one using an unsupervised harmonization technique (Liu et al., 2021). The reason for using an unsupervised technique is that none of our datasets provide the ground truth of the same MRI with different scanners. This model leverages Generative Adversarial Networks (GANs) for MRI harmonization. Adapting the style of MRI images from one site to match the style of another, reduces inter-site variability and improves the generalizability of MRI-based models. Using this unsupervised harmonization technique, we harmonized all of our data. We then incorporated these harmonized data into both the CDSSL framework, which we refer to as "Harmonized-Data CDSSL", and into the fine-tuning process, where we use the labels to predict the Clinical Dementia Rating Sum of Boxes (CDR-SB) score using baseline MRIs.

Figure 11 illustrates a workflow for integrating harmonization models with self-supervised learning techniques to enhance medical imaging tasks. The process begins with both unlabeled and labeled medical images being input into a harmonization model to standardize the images. Harmonized unlabeled images are then fed into a self-supervised model with a CNN backbone architecture, which is trained using contrastive loss to learn meaningful representations. Simultaneously, harmonized labeled images undergo knowledge transfer from the self-supervised model's CNN backbone. This pretrained backbone is then fine-tuned on labeled data for specific downstream tasks, leveraging the learned representations to improve performance on these tasks. This approach combines the benefits of harmonization for image standardization and self-supervised learning for representation learning, ultimately enhancing the effectiveness of medical image analysis.

3.4 Experimental Study: Cross-Domain Self-Supervised Learning

To evaluate the effectiveness of different pretraining models for disease progression prediction, we proposed the first benchmarking study. This study explores the transferability

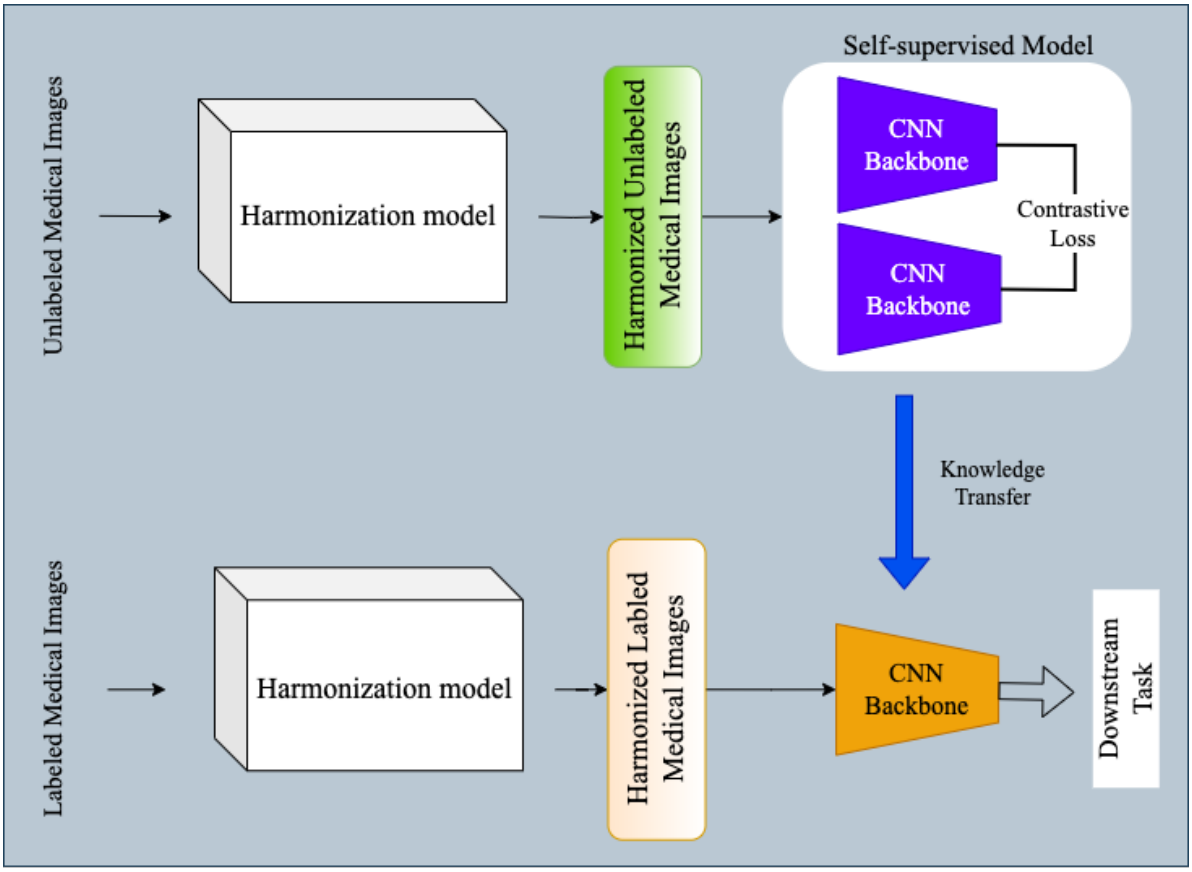


Figure 11: This figure illustrates Harmonized-Data CDSSL which integrates harmonization and SSL for enhancing prognosis models.

of features learned by pretraining on natural or medical images, or both, to the medical task of predicting disease progression. To evaluate our model's performance, we used the Pearson correlation coefficient (r) and the coefficient of determination (R^2). R^2 is calculated as equation 2:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{2}$$

Where $\sum (y_i - \hat{y})^2$ and $\sum (y_i - \bar{y})^2$ respectively indicate the sum of squared residuals and total sum squared. In clinical studies, R^2 is widely employed to evaluate the ability of a model to predict future outcomes (Franzmeier et al., 2020).

3.4.1 Self-supervised pretraining

Experiments on natural images. Three widely known SSL methods were used to test the transferability of standard ImageNet models: SimCLR, Barlow Twins, and SwAV. An ImageNet dataset is used for pretraining, and a ResNet-50 backbone is used in all SSL models.

Results. The results in Table 4a show that transfer learning from the supervised ImageNet model does not improve over random initialization. The reasons for this may be due to the significant difference between pretraining tasks and regression targets. A supervised ImageNet model, for example, may capture semantic features that are specific to a particular domain. Therefore, it is inefficient to use supervised imagenet models if the data distributions in the target and pretraining datasets are different. The results of this study are consistent with studies on other medical tasks that found transfer learning from supervised ImageNet pretraining did not always correlate with performance on classification or segmentation (Dippel et al., 2021; Vendrow and Schonfeld, 2022; Hosseinzadeh Taher et al., 2021).

In contrast, transfer learning from self-supervised ImageNet models provides superior performance compared with both random initialization and transfer learning from the supervised ImageNet model. The best self-supervised model (i.e., Barlow Twins) achieves a performance improvement of 7% and 8% over random initialization and the supervised ImageNet model, respectively. This is likely due to the fact that self-supervised ImageNet models are trained to learn general-purpose features that are not biased toward any particular task. As a result, they are better able to generalize to new domains.

Experiments on medical images. To investigate the effect of using in-domain medical images for self-supervised pretraining, we trained three SSL methods, SimCLR, Barlow Twins, and SwAV, on 5 unlabeled medical imaging datasets, which we call in-domain datasets. All SSL models were randomly initialized and then fine-tuned on our labeled dataset.

Results. Table 4b shows the performance of SSL models pretrained on the 5-slice dataset, measured by the R^2 score. We observe that SimCLR pretraining on the in-domain

Table 4: The effects of different pretraining schemes on downstream tasks.

(a) Pretraining on natural images.

Pretraining	Initialization	R^2
-	Random	0.07
Supervised	ImageNet	0.06
Self-Supervised	SimCLR	0.10
	SwAV	0.08
	Barlow Twins	**0.14**

(b) Pretraining on medical images.

Pretraining	Initialization	R^2
-	Random	0.07
Self-Supervised	SimCLR	0.19
	SwAV	0.12
	Barlow Twins	0.14

dataset achieves the highest performance, providing a 7% and 5% boost over SwAV and Barlow Twins, respectively. This may be due to the superiority of contrastive learning for identifying significant MRI features for predicting the progression of Alzheimer's disease in terms of CDR-SB. Moreover, the performance of SimCLR pretraining on the in-domain dataset exceeds that of both supervised and self-supervised pretraining on the ImageNet dataset (as seen in Table 4a). This suggests that pretraining on the in-domain dataset encodes domain-specific features that reflect the distinctive characteristics of medical images.

In contrast, pretraining Barlow Twins on in-domain data does not yield performance improvement compared to Barlow Twins pretrained on ImageNet. This result indicates that the features learned by Barlow Twins through pretraining on ImageNet demonstrate sufficient generalizability to medical images. Thus, the limited number of unlabeled medical images in the in-domain dataset (40k compared to 1.3M in ImageNet) may only provide marginal performance gains for the redundancy reductions-based Barlow Twins method.

Cross-domain experiments. In this experiment, we examine the effects of self-supervised pretraining on both natural images and medical images. We achieved this by pretraining SimCLR on the five-slice dataset using two distinct initialization schemes: Supervised ImageNet (referred to as ImageNet(Labeled)→In-domain), and Barlow Twins on ImageNet (referred to as ImageNet(Unlabeled)→In-domain). In our experiments, SimCLR

and Barlow Twins were selected because their performance was the highest when pretrained on natural or medical images, respectively, as shown in Tables 4a and 4b. Figure 8(b,c) shows both cross-domain self-supervised pretraining schemes. As part of this section, we also include Pearson correlation coefficients to illustrate the strength and direction of the relationship between CDR-SB predicted values and target values.

Results. The results are displayed in Table 5a. When both unlabeled ImageNet and in-domain datasets are used to pretrain, the best results are achieved. Specifically, the predictive power of the ImageNet(Unlabeled)→In-domain model outperforms that of the model trained only on the ImageNet or the in-domain dataset. The performance improvements are 14%, 7%, and 2% when compared to random initialization, ImageNet pretraining alone, and in-domain dataset pretraining alone. These results indicate that pretraining on ImageNet combined with pretraining on in-domain datasets leads to more robust representations for medical applications, as suggested by earlier research (Hosseinzadeh Taher et al., 2021; Azizi et al., 2021). A further point worth emphasizing is that the ImageNet(Labeled)→In-domain pretraining method shows inferior performance compared to ImagesNet(Unlabeled)→In-domain methods. In these observations, self-supervised models demonstrate their effectiveness at generating more generic representations that can be applied to target tasks with limited data, reducing the need for extensive annotations.

To compare the Pearson correlation coefficients of the models in Table 5a, we used Steiger's Z1 method (Steiger, 1980). This method is utilized to compute the two-tailed p-value at a 95% confidence interval, and it's a statistical approach uniquely suited for assessing the significance of variances between dependent correlation coefficients. The results are displayed in Table 6. Notably, the ImageNet(Unlabeled)→In-domain pretraining model demonstrates a statistically significant difference compared to other models.

3.4.2 In- and out-of-domain generalization

To assess the robustness of our top-performing model, Barlow Twins→SimCLR, on the 5-slice dataset, we evaluated its performance using three distinct test sets: an in-study test set and two out-study test sets. Furthermore, we compare this model's performance with

Table 5: Results of different pretraining schemes on both (a) validation and (b) test sets in terms of R^2 and r. OASIS-3 is an out-study test set, meaning it has not been utilized for either SSL pretraining or fine-tuning.

(a) Results of the top-performing models in each domain and their combination in a cross-domain SSL setting on the validation set.

Pretraining Method	Pretraining Dataset	R^2	r	MSE
Random	-	0.07	0.33	5.31
Barlow Twins	ImageNet	0.16	0.42	4.81
SimCLR	In-domain	0.19	0.44	4.61
Supervised ImageNet → SimCLR	ImageNet(Labeled) → In-domain	0.17	0.43	4.74
Barlow Twins → SimCLR	ImageNet(Unlabeled) → In-domain	**0.21**	**0.46**	**4.52**

(b) Results on independent dataset.

Pretraining Method	Pretraining Dataset	R^2/r on in-study test	R^2/r on OASIS-3
Random	-	0.04/0.30	-0.04/0.10
Barlow Twins	ImageNet	0.12/0.35	-0.06/0.15
SimCLR	In-domain	0.14/0.38	0.11/0.36
Supervised ImageNet → SimCLR	ImageNet(Labeled) → In-domain	0.11/0.33	0.10/0.35
Barlow Twins → SimCLR	ImageNet(Unlabeled) → In-domain	**0.18/0.42**	**0.17/0.42**

the best-performing models initialized either by ImageNet or an in-domain dataset.

According to the results presented in Table 5b, the highest performance is achieved when both the unlabeled ImageNet and in-domain dataset are utilized for pretraining. Specifically, the ImageNet(Unlabeled)→In-domain pretraining approach exhibits a significant improvement of 10% and 6% over the in-domain and ImageNet pretrained models, respectively, for the in-study test set. Similarly, on the out-study test set OASIS-3, the same model demonstrates an improvement up to 6% compared to other models. These results indicate that

Table 6: The p-values in Table 5a show the statistical significance of the difference between the Pearson Correlation (r) of different models. Significance levels are denoted by *, **, and *** for p-values < 0.05, < 0.01, and < 0.001, respectively.

Pretraining Method	Random	Barlow Twins	SimCLR	Supervised ImageNet → SimCLR	Barlow Twins → SimCLR
Random	-	0.003**	0.0001***	0.0001***	0.0001***
Barlow Twins	0.003**	-	0.04*	0.27	0.0004***
SimCLR	0.0001***	0.04*	-	0.35	0.012*
Supervised ImageNet → SimCLR	0.0001***	0.27	0.35	-	0.004**
Barlow Twins → SimCLR	0.0001***	0.0004***	0.012*	0.004*	-

ImageNet(Unlabeled)→In-domain pretraining effectively encodes semantic features that are generalizable to other studies. Furthermore, the performance of ImageNet(Labeled)→In-domain pretraining is inferior to that of ImageNet(Unlabeled)→In-domain pretraining. This indicates that supervised ImageNet models encode domain-specific semantic features, which may not be efficient when the pretraining and target data distributions significantly differ.

3.4.3 Visualizing model saliency maps

Attribution methods are a tool for investigating and validating machine learning models. Using the interpretability of the ML models can significantly help obtaining a bigger picture about risk factors influences on short-term prognosis. We used the GradCAM (Selvaraju et al., 2017) method to extract and evaluate the varying importance of each part of brain MRIs using a gradient of the final score. In this method, regions of an image are marked with different colors ranging from red to blue. Generally, areas that are closer to the red color contribute more significantly to the final result, based on the input data (i.e., MRI slice).

Figure 12 presents various examples of saliency maps, which depict the significance of different regions in the MRIs at a pixel level. Notably, our top-performing model, Barlow Twins→SimCLR, consistently highlights the subcortical areas of the brain. This observation

aligns with prior research indicating that the initial stages of AD exhibit abnormal tau accumulation in the entorhinal cortex and subcortical brain regions (Rueb et al., 2017; Liu et al., 2012). Therefore, it is reasonable that our model, i.e. Barlow Twins→SimCLR, exhibits higher attention to those specific brain regions in a population of individuals with prodromal to mild Alzheimer's disease at baseline. As expected, the randomly initialized model highlights random features all around MRIs, including background areas. In contrast, the model initialized with unsupervised ImageNet is more focused on brain regions, rather than irrelevant areas such as the background. Table 7 illustrates the results of the Cross-Domain

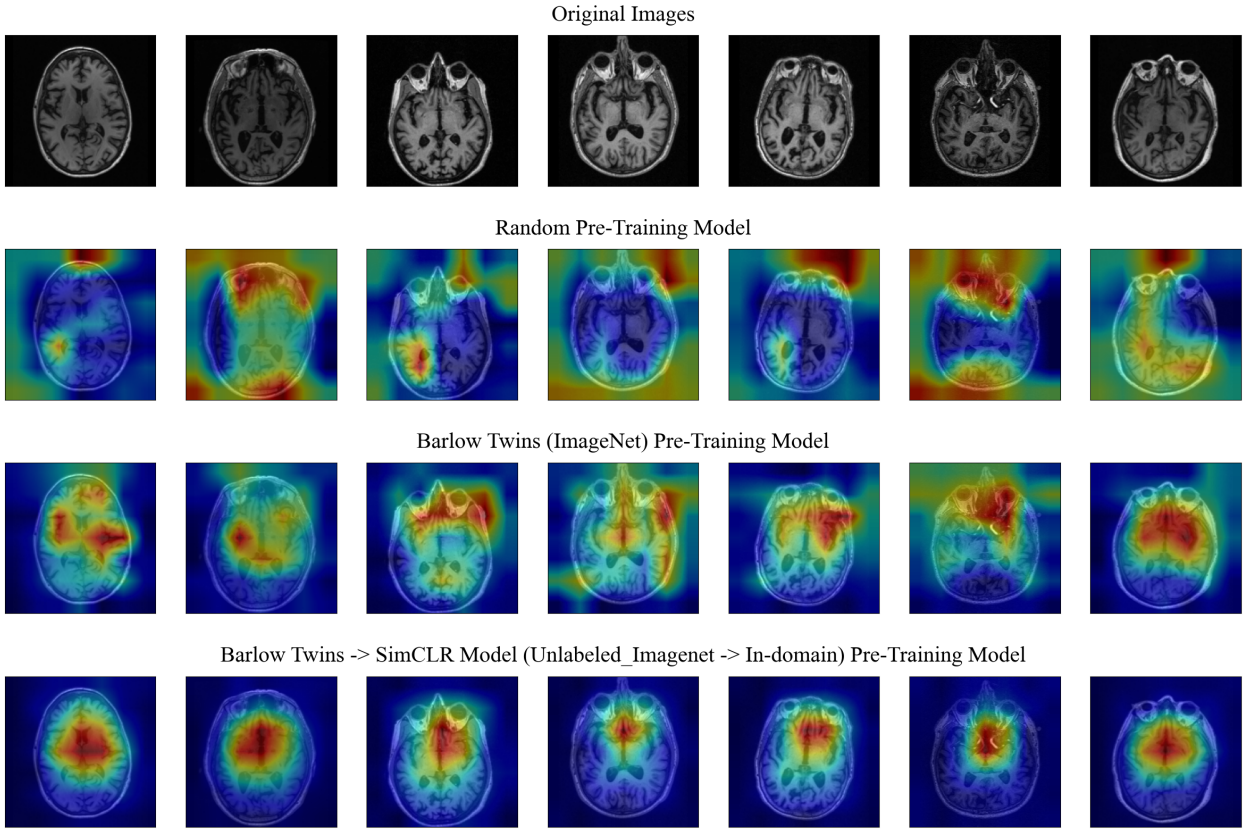


Figure 12: This figure illustrates the interpretation of three pretraining models using the GradCam technique. The top row shows the original MRI slices, while the subsequent rows depict the saliency maps generated by the following models: a randomly initialized pretrained model, a pretrained model on natural images, and our best model, Barlow Twins→SimCLR.

Self-Supervised Learning (CDSSL) framework, focusing on the percentage of coverage in

clinically relevant brain areas—ventricles, hippocampus, and amygdala—by heatmaps using three pre-training strategies. These areas are critical in the early stages of Alzheimer's disease. Numeric calculations show the percentage of coverage, revealing significant differences across the pre-training models. In the random pre-training model, the heatmap areas are scattered randomly, resulting in a low average percentage of coverage for each clinically related area: 22% for ventricles, 30% for the hippocampus, and 20% for the amygdala. The mean and standard deviation (Std) values reflect this inconsistency. With one-domain self-supervised learning (SSL) as a pre-training strategy, there is an improvement in these coverage percentages: 50% for ventricles, 60% for the hippocampus, and 61% for the amygdala, showing more focused heatmaps. The cross-domain SSL model further enhances these results, with mean coverage percentages of 95% for ventricles, 89% for the hippocampus, and 97% for the amygdala. The heatmaps are more consistent across subjects, indicated by lower standard deviations. The high mean coverage and low standard deviation for the amygdala are due to its small size in some 2D slices, where it might not be present, leading to 0% coverage for those slices. This absence contributes to the low standard deviation compared to the mean.

Table 7: Heatmap coverage of clinically relevant brain areas using different pretraining strategies.

Pretraining Methods	Pretraining Dataset	Ventricles mean%(std)	Hippocampus mean%(std)	Amygdala mean%(std)
Random	-	22% (21)	50%(18)	95%(3)
Barlow Twins	ImageNet(Unlabeled)	30%(31)	60%(13)	89%(2)
Barlow Twins → SimCLR	ImageNet(Unlabeled) → In-domain	20%(18)	61%(10)	97%(10)

3.4.4 Evaluating the addition of CDR-SB at baseline with the baseline MRI

The hypothesis of this study is that adding clinical baseline information, specifically CDR-SB score, to our supervised model will improve prediction results. The rationale behind this hypothesis is that using baseline MRI images in conjunction with the CDR-SB

score allows the model to better learn the correlation between the baseline image and the progression prediction of the patient. This enhanced learning is expected to improve the model's ability to predict future clinical outcomes. To test this hypothesis, we incorporated the CDR-SB score at the baseline into our model. This approach involves using the CDR-SB score as an additional input feature alongside the baseline MRI images. The goal is to determine how much the inclusion of this clinical baseline information improves the model's predictive performance.

Results. The results of this strategy are illustrated in the following Figure 13. The left graph shows the Pearson correlation coefficient between the predicted and actual CDR-SB scores, while the right graph shows the coefficient of determination for the same predictions.

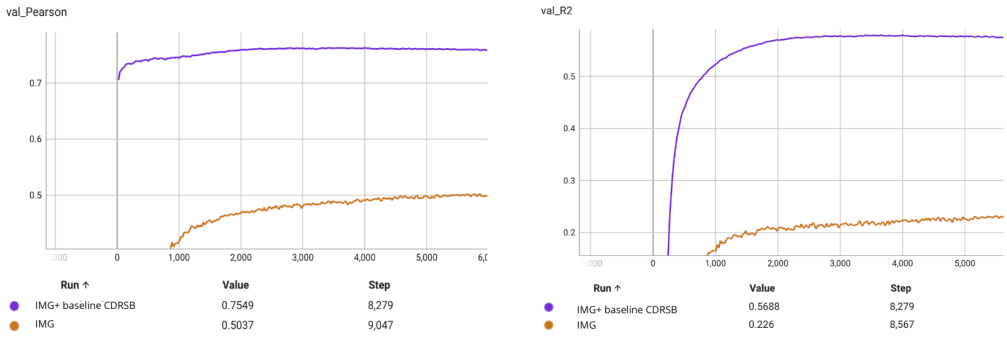


Figure 13: Plots of adding CDR-SB baseline information to the pipeline.

In both graphs, the purple line represents the model that includes the baseline CDR-SB score (MRI + baseline CDR-SB), and the orange line represents the model that uses only the baseline MRI images. The results clearly indicate that the model incorporating the baseline CDR-SB score outperforms the model using only MRI images. Specifically, the Pearson correlation coefficient for the MRI+ baseline CDR-SB model reaches 0.7549, compared to 0.5037 for the MRI model. Similarly, the coefficient of determination (R2) for the MRI + baseline CDR-SB model is 0.5688, compared to 0.226 for the MRI model. These findings suggest that adding the CDR-SB score at the baseline significantly enhances the model's ability to predict Alzheimer's disease progression, demonstrating the value of incorporating clinical baseline information into predictive models.

3.5 Using SSL to Address Scanner Effect

3.5.1 Evaluating the efficacy of CDSSL in reducing scanner effect

In this section, we use a supervised setting to find the efficacy of our CDSSL using R^2, Pearson correlation (r), mean squared error, and other relevant metrics against other baseline models such as only supervised ImageNet, only unsupervised ImageNet, Random initialization, and the model that we incorporate scanner information including manufacturer and model. Furthermore, when predicting imaging-derived scores like brain volume, metrics such as the correlation coefficient between predicted and actual values, or the mean absolute error, can offer additional evaluation dimensions.

In order to find whether or not transfer learning using different methods such as CDSSL or only ImageNet, or random initialization affected by scanner variation in predicting clinical and imaging score, we compare two nested models represented in Figure9 that fuse the scanner manufacturer and scanner model to the supervised framework embedding. To discern the performance difference between two nested deep learning models, a Likelihood Ratio Test (LRT) was employed. Initially, the log-likelihood of both models was computed on the validation set. The LRT statistic was then determined by doubling the difference between the log-likelihoods of the two models. Subsequently, the degrees of freedom, essential for the chi-squared distribution, were derived from the discrepancy in the number of parameters between the models. With the LRT statistic and the degrees of freedom, the associated p-value was calculated. This p-value offers critical insight: if significantly low, it suggests that Model 2 fits the data more effectively than Model 1.

Results. As an initial step, we only choose labeled data from ADNI datasets. This resulted in 142 MRIs for training and 68 MRIs for validation. As can be seen in Table 8 distinct observations were discerned. Under the Barlow Twins → SimCLR pretraining with only MRI data, a Pearson correlation of 0.39 was observed. Incorporating manufacturer data resulted in a slight decrease in correlation to 0.34. Adding further, with model information, led to a correlation value of 0.40. In the ImageNet pretraining scenario, using only MRI data yielded a correlation of 0.25. Remarkably, the inclusion of both manufacturer and model

data did not change this correlation, maintaining it at 0.25. For the Random pretraining environment, the MRI model showed a correlation of 0.19. However, when both manufacturer and model data were added, there was a notable increase in the correlation, reaching 0.33. Turning to the LR test results, which are crucial for assessing model fit differences, several insights emerged. As can be seen in Table 9 within the Barlow Twins → SimCLR context, the comparison of various models, ranging from MRI to those incorporating manufacturer and model information, consistently yielded p-values of 1, suggesting no significant differences in model fit. A congruent trend was found in the ImageNet pretraining, with a stable p-value of 1. The narrative shifted dramatically in the Random pretraining. Here, contrasting the MRI model with the combined manufacturer and model data led to a significant LR statistic of -164, indicating a profound difference in model fit with a p-value of less than 0.001.

In conclusion, while the Barlow Twins → SimCLR and ImageNet pretraining datasets indicated stable model fits irrespective of the incorporated data, the Random pretraining setting demonstrated substantial variability. Within this latter regime, the integration of both manufacturer and model details alongside MRI data showed a pronounced enhancement in model fit.

Table 8: Comparison of r and log-likelihood values across different pretraining techniques (Barlow Twins → SimCLR, Supervised ImageNet, and Random) for MRI and combined data settings (MRI with manufacturer and MRI with manufacturer and model).

Pretraining	Dataset	Pearson Correlation (r)	Log-likelihood
Random	MRI	0.19	-445
	MRI + Manufacturer +Model	0.33	-363
Supervised ImageNet	MRI	0.25	-430
	MRI + Manufacturer +Model	0.25	-463
Barlow Twins → SimCLR	MRI	0.39	-384
	MRI + Manufacturer	0.34	-395
	MRI + Manufacturer +Model	0.40	-397

Table 9: Statistical significance of the resulting likelihood ratio. The p-values indicate the statistical significance of the difference between the log-likelihood of different inputs for each pre-training using a degree of freedom (df). Significance levels indicated by *, **, and *** for p-values < 0.05, < 0.01, and < 0.001, respectively.

		Likelihood Ratio	
Pretraining	Datasets	MRI + Manufacturer	MRI + Manufacturer+Model
Barlow Twins → SimCLR	MRI	-26(df=4) p-value =1	-21(df=38) p-value =1
	MRI + Manufacturer	-	5(df=34) p-value =1
Supervised ImageNet	MRI	-	-64(df=38) p-value =1
Random	MRI	-	164(df=38) **p-value <0.001 *****

3.5.2 Enhancing CDSSL with Scanner Information Incorporation

All the evaluations in this section will be done on both self-supervised and supervised frameworks. Within the self-supervised context, the quality of the embeddings can be gauged by analyzing the distribution within the learned feature space. Visualization techniques like t-SNE or UMAP can offer insights into the compactness and separability of these embeddings, spotlighting the model's adeptness in capturing both scanner-specific nuances and underlying biological variations.

Transitioning to the supervised setting, the utility of the model becomes more pronounced when evaluating its performance in predicting specific clinical and imaging scores. By fine-tuning the pre-trained model on tasks related to predicting clinical scores, such as CDR-SB one can benchmark its predictive performance.

Results. According to Table 10, Auxiliary CDSSL shows a performance improvement when scanner information is included. The R^2 value increases from 0.62 to 0.64, and the cor-

relation coefficient (r) increases slightly from 0.8414 to 0.8454. This indicates that including scanner information helps in better capturing the variance and improving the correlation with the target variables. Similar to Auxiliary CDSSL, Augmentation CDSSL also benefits from the inclusion of scanner information. The R^2 value improves from 0.63 to 0.645, and the correlation coefficient (r) remains relatively stable at around 0.8445. This suggests that augmenting the dataset with scanner information provides a small but noticeable improvement in model performance. The Original CDSSL method sees a modest increase in performance with the addition of scanner information. The R^2 value increases from 0.55 to 0.57, while the correlation coefficient (r) remains constant at 0.83. This indicates that while there is some benefit to adding scanner information, it is less pronounced compared to the other CDSSL methods. For the ImageNet method, the performance slightly decreases when scanner information is included. The R^2 value drops from 0.20 to 0.19, and the correlation coefficient (r) decreases from 0.82 to 0.81. This suggests that for this particular method, scanner information may introduce noise or irrelevant information that negatively impacts performance. The Random method serves as a baseline, and as expected, it has the lowest performance metrics. The R^2 value slightly increases from 0.14 to 0.16, and the correlation coefficient (r) improves from 0.71 to 0.74 when scanner information is included. Despite these improvements, the overall performance remains low, highlighting the effectiveness of the other methods.

Table 10: Comparison of different methods on the ADNI dataset with performance metrics R^2 and r for MRI and MRI with Scanner information conditions.

Pretraining Method	MRI	MRI + Scanner Info	p-value (LR)
Auxiliary CDSSL	$R^2 = 0.62$, r = 0.84	$R^2 = 0.64$, r = 0.84	0.91
Augmentation CDSSL	$R^2 = 0.63$, r = 0.84	$R^2 = 0.645$, r = 0.84	1
Original CDSSL	$R^2 = 0.55$, r = 0.83	$R^2 = 0.57$, r = 0.83	0.95
ImageNet	$R^2 = 0.20$, r = 0.82	$R^2 = 0.19$, r = 0.81	1
Random	$R^2 = 0.14$, r = 0.71	$R^2 = 0.16$, r = 0.74	0.85

3.5.2.1 Performance comparison of pretraining methods across scanner manufacturers in the ADNI dataset

Table 11 presents the average Mean Squared Error (MSE) for different pretraining methods across three major scanner manufacturers in the ADNI dataset: Siemens, GE, and Philips.

Table 11: Average MSE using different pretraining methods for each scanner manufacturer in the ADNI dataset.

Pretraining Method	Siemens	GE	Philips
Augmentation CDSSL	1.66	1.15	1.72
Auxiliary CDSSL	1.67	1.17	1.79
Original CDSSL	1.72	1.25	1.86
ImageNet	2.22	1.85	2.58
Random	2.38	1.83	2.45

Augmentation CDSSL and Auxiliary CDSSL methods show comparable performance across all manufacturers, with slightly lower MSEs for Siemens and GE compared to Philips. Specifically, Augmentation has MSE values of 1.66 for Siemens, 1.15 for GE, and 1.72 for Philips, while Auxiliary has MSEs of 1.67, 1.17, and 1.79 respectively.

The Original CDSSL method has slightly higher MSE values than both Augmentation and Auxiliary, with Siemens at 1.72, GE at 1.25, and Philips at 1.86, indicating that augmentation and auxiliary techniques may provide a slight advantage in reducing errors.

ImageNet pretraining results in higher MSEs across all manufacturers compared to the previously mentioned methods, with values of 2.22 for Siemens, 1.85 for GE, and 2.58 for Philips, suggesting that it may be less effective for this specific dataset and task.

The Random initialization method has the highest MSE values for Siemens and Philips, at 2.38 and 2.45 respectively, while it performs similarly to ImageNet for GE with an MSE of 1.83. This highlights the importance of using more sophisticated pretraining methods over

random initialization for improving model performance.

Overall, Augmentation and Auxiliary methods appear to be the most effective in reducing MSE for the ADNI dataset across different scanner manufacturers, with Original also performing relatively well. ImageNet and Random methods are less effective, particularly for Siemens and Philips scanners.

3.5.2.2 Clustering performance for ADNI dataset

In a well-executed feature harmonization process, the objective is to reduce the variability introduced by different scanners, ensuring that the features from each scanner are less clustered and more homogeneous. This means that after harmonization, data points from different scanners should be indistinguishable from one another in the feature space, reflecting true biological or experimental variations rather than technical artifacts. To evaluate the effectiveness of this harmonization, metrics such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index can be employed. A successful harmonization will result in a lower Silhouette Score, indicating that the data points are less tightly clustered by scanner, and a lower DBI, suggesting reduced intra-scanner similarity. Additionally, a higher CHI will indicate that the clusters formed are more defined by biological differences rather than scanner differences. These metrics collectively help in quantifying and demonstrating the reduction of scanner-induced clustering, affirming the effectiveness of the harmonization process.

The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a higher score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. The Silhouette Score is useful for understanding the cohesion and separation of clusters.

The Silhouette Score for sample i is:

$$Silhouette(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3}$$

Equation 3 shows this score for a sample i: $a(i)$ is the average distance between i and all other points in the same cluster. $b(i)$ is the average distance between i and all points in the nearest cluster. The overall Silhouette Score is the mean of $Silhouette(i)$ for all samples.

The **Davies-Bouldin Index (DBI)** evaluates the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DBI indicates better clustering, with well-separated clusters that are internally compact. DBI is useful for comparing different clustering algorithms or configurations on the same dataset.

The DBI is:

$$DBI = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{4}$$

Equation 4 shows this index for n clusters, C_i and C_j. S_i is the average distance between each point in cluster i and the centroid of cluster i. M_{ij} is the distance between the centroids of clusters i and j.

The **Calinski-Harabasz Index (CHI)**, also known as the Variance Ratio Criterion, measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion for all clusters. A higher CHI indicates better-defined clusters. This metric is effective for evaluating the overall goodness of fit of clustering results.

The CHI is:

$$CHI = \frac{\text{trace}(B_k)}{\text{trace}(W_k)} \cdot \frac{(n-k)}{(k-1)} \tag{5}$$

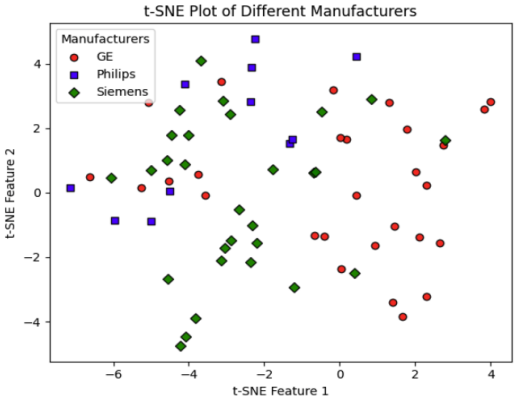
Equation 5 demonstrates this index for n samples, k clusters. B_k is the between-cluster dispersion matrix, calculated as the sum of squared differences between the cluster centroids and the overall centroid, weighted by the number of points in each cluster. W_k is the within-cluster dispersion matrix, calculated as the sum of squared differences between each point and its respective cluster centroid.

All these metrics provide different perspectives on the quality of clustering results, making them useful for comprehensive cluster analysis.

Results. Tables 12 and 13 present the clustering performance metrics before and after fine-tuning a supervised model for CDR-SB prediction for each pertaining method. Table 12 shows the metrics derived from features obtained through Augmentation CDSSL, while Table 13 shows the metrics after fine-tuning the model for CDR-SB prediction.

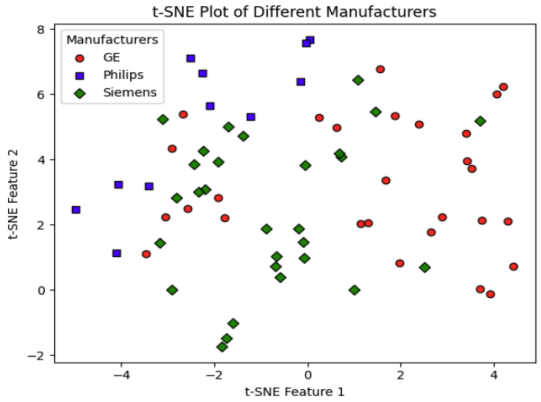
The low Silhouette Score (0.03) in Table 12 suggests that the clustering structure is weak, indicating that the features are scattered more across different manufacturers. The

Table 12: Clustering performance metrics - before fine-tuning (Augmentation CDSSL).



Metric	Value
Silhouette Score	0.03
DBI	2.51
CHI	7.42

Table 13: Clustering performance metrics - after fine-tuning (CDR-SB prediction model).



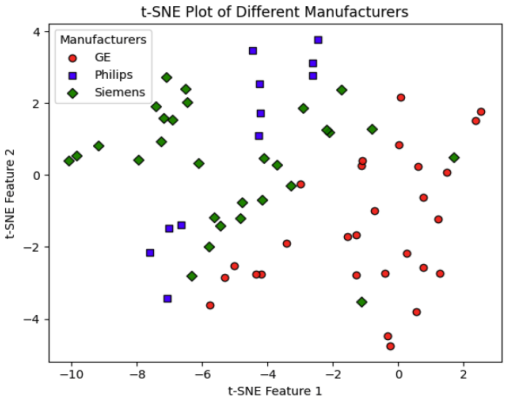
Metric	Value
Silhouette Score	0.07
DBI	2.14
CHI	10.02

DBI (2.51) is relatively high, which further confirms that the clusters are not compact and well-separated. The CHI (7.42) is also low, reflecting poor clustering performance. After fine-tuning the model for CDR-SB prediction, there are noticeable improvements in the clustering performance metrics. In Table 13, the Silhouette Score increases to 0.07, suggesting a slight improvement in the clustering structure. The DBI decreases to 2.14, indicating better compactness and separation of clusters. The CHI significantly increases to 10.02, highlighting a more favorable clustering structure. The comparison between the two tables highlights the impact of fine-tuning the model for CDR-SB prediction. The clustering performance metrics improve across all three indices:

Table 14 shows the metrics derived from features obtained through Auxiliary CDSSL, while Table 15 shows the metrics after fine-tuning the model for CDR-SB prediction.

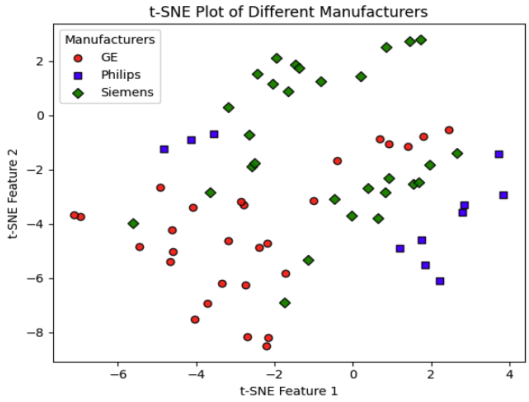
The Silhouette Score of 0.13 indicates that the clustering structure is weak, but not

Table 14: Clustering performance metrics - before fine-tuning (Auxiliary CDSSL).



Metric	Value
Silhouette Score	0.13
DBI	6.48
CHI	17.67

Table 15: Clustering performance metrics - after fine-tuning (CDR-SB prediction model).

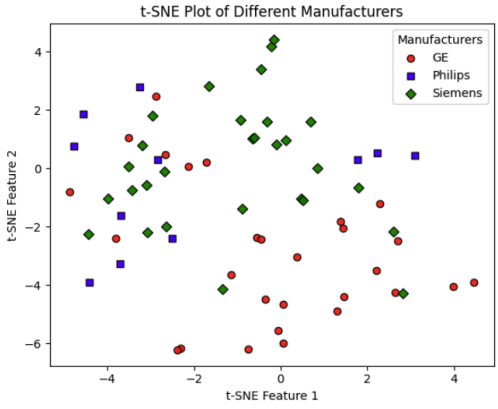


Metric	Value
Silhouette Score	0.13
DBI	2.25
CHI	10.94

as poor as some other methods. The DBI of 6.48 is relatively high, suggesting that the clusters are not very compact or well-separated. The CHI of 17.67 is higher than some other initial metrics, indicating a somewhat better clustering structure before fine-tuning. After fine-tuning the model for CDR-SB prediction, the Silhouette Score remains the same at 0.13, indicating that the overall cohesion and separation of clusters have not significantly changed. However, the DBI decreases significantly to 2.25, indicating improved compactness and separation of clusters. The CHI decreases to 10.94, which, while lower than before, still suggests a reasonable clustering structure. Overall, the fine-tuning process for the Auxiliary CDSSL method has led to more compact and better-separated clusters, as evidenced by the significant improvement in the DBI, even though the Silhouette Score and CHI present a more nuanced view. This suggests that fine-tuning helps achieve more well-defined clusters that are easier to interpret and use for subsequent predictive tasks.

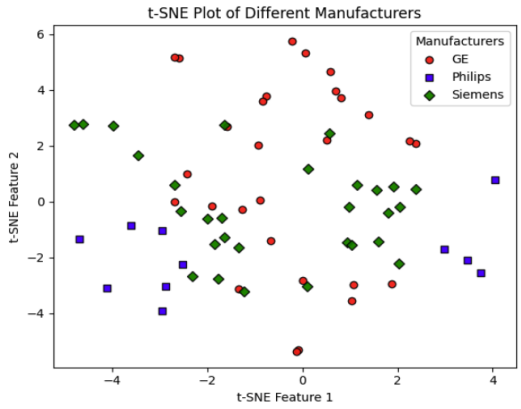
Table 16 shows the metrics derived from features obtained through Original CDSSL, while Table 17 shows the metrics after fine-tuning the model for CDR-SB prediction.

Table 16: Clustering performance metrics - before fine-tuning (Original CDSSL).



Metric	Value
Silhouette Score	0.06
DBI	4.19
CHI	8.16

Table 17: Clustering performance metrics - after fine-tuning (CDR-SB prediction model).



Metric	Value
Silhouette Score	0.01
DBI	4.25
CHI	3.28

The Silhouette Score of 0.06 indicates a weak clustering structure, suggesting that the clusters are not well-separated. The DBI of 4.19 is relatively high, which indicates that the clusters are not very compact or well-separated. The CHI of 8.16 reflects moderate clustering performance before fine-tuning. After fine-tuning the model for CDR-SB prediction, the Silhouette Score decreases to 0.01, indicating a further weakening of the clustering structure. The DBI increases slightly to 4.25, suggesting that the clusters have become less compact and less well-separated. The CHI decreases significantly to 3.28, indicating a deterioration in the clustering performance. Overall, the fine-tuning process for the Original CDSSL method has not led to improvements in the clustering structure. In fact, the metrics suggest that the clusters have become less distinct and more poorly defined after fine-tuning. This highlights the potential challenges of using the Original CDSSL method for this specific task

and dataset, suggesting that alternative pretraining methods or further optimization may be needed to achieve better clustering performance and subsequent predictive accuracy.

The provided Table 18 presents the clustering and prediction metrics for various pretraining methods before and after fine-tuning. The goal is to evaluate these methods based on their ability to achieve higher R^2 prediction power and lower clustering metrics, which indicate better harmonization across different scanners. The harmonic mean composite scores for each method after fine-tuning are also provided, offering an integrated measure of overall performance.

Auxiliary CDSSL (After Fine Tuning) has the highest harmonic mean composite score of 0.41, suggesting that it achieves the best balance of prediction power and clustering performance. The high R^2 value of 0.62 indicates strong predictive power, and the low DBI of 2.25 shows improved cluster compactness and separation. Augmentation CDSSL (After Fine Tuning) also shows good performance with a harmonic mean composite score of 0.24. It has the highest R^2 of 0.63 but slightly higher clustering metrics compared to Auxiliary CDSSL, indicating somewhat less optimal clustering despite strong prediction power.

Table 18: Clustering and prediction metrics before and after fine-tuning.

Method	Silhouette Score	DBI	CHI	R^2 (Prediction Power)
Auxiliary CDSSL (Before Fine-Tuning)	0.13	6.48	17.67	-
Auxiliary CDSSL (After Fine-Tuning)	0.13	2.25	10.94	0.62
Augmentation CDSSL (Before Fine-Tuning)	0.03	2.51	7.42	-
Augmentation CDSSL (After Fine-Tuning)	0.07	2.14	10.02	0.63
Original CDSSL (Before Fine-Tuning)	0.06	4.19	8.16	-
Original CDSSL (After Fine-Tuning)	0.01	4.25	3.28	0.57
ImageNet (After Fine-Tuning)	0.02	3.26	5.23	0.19
Random (After Fine-Tuning)	0.05	2.74	10.21	0.16

Original CDSSL (after fine-tuning) has a low harmonic mean composite score of 0.04. Despite a decent R^2 of 0.57, the high DBI of 4.25 and low CHI of 3.28 indicate poor clustering

performance. ImageNet (after fine-tuning) has a lower harmonic mean composite score of 0.07. The low R^2 of 0.19 indicates poor predictive power, and the clustering metrics suggest suboptimal cluster quality. Random (after fine-tuning) yields a moderate harmonic mean composite score of 0.15. The clustering metrics are reasonable, but the low $\mathbf{R^2}$ of 0.16 indicates weak predictive power.

In conclusion, the harmonic mean composite scores suggest that Auxiliary CDSSL (after fine-tuning) is the best pretraining method, achieving the highest overall performance by balancing prediction power and clustering quality. Augmentation CDSSL (after fine-tuning) also performs well, particularly in terms of prediction power, but with slightly less optimal clustering metrics. Other methods, including Original CDSSL, ImageNet, and Random, show lower overall performance, indicating less effective pretraining for the given tasks.

3.5.3 Comparative Analysis with Unsupervised Harmonization Techniques

In this section, we want to compare the efficacy of CDSSL and unsupervised harmonization techniques. Therefore, using the method proposed by (Liu et al., 2021), we choose one scanner/style and harmonized all the images into that same scanner/style. Then, similar to the previous section, we apply the images in a supervised setting and compute scores including R^2 and Pearson correlation (r) and also the log-likelihood ratio of how statistically significant these methods predict the clinical scores.

Table 19: Performance of different pretraining methods on various harmonized datasets.

Pretraining	Random Initializer	Original CDSSL	Harmonized-Data CDSSL
Harmonized-ADNI	0.05	0.40	0.63

Table 20: Performance of different pretraining methods on various datasets.

Pretraining	Random Initializer	Original CDSSL	Augmentation CDSSL	Auxiliary CDSSL
ADNI	0.14	0.20	0.63	0.62

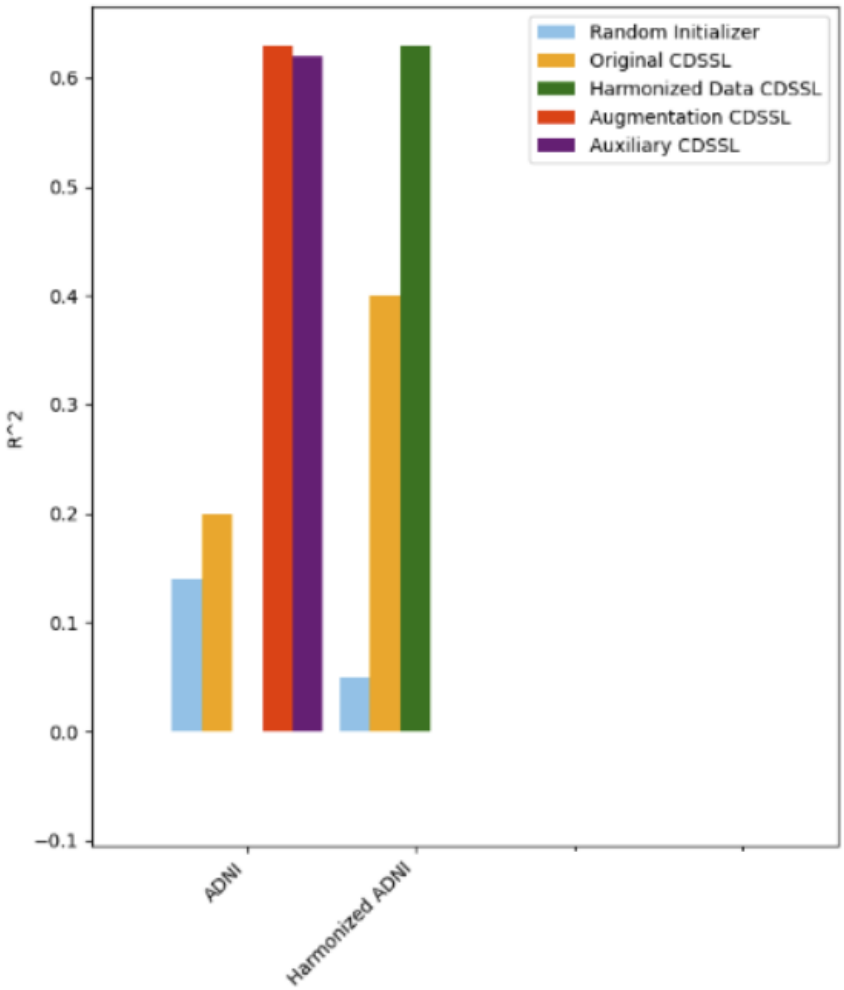


Figure 14: Comparison of pretraining methods across original ADNI dataset and Harmonized-ADNI.

The tables and combined plots illustrate the performance of various pretraining meth-

ods across different datasets. Table 19 focuses on harmonized datasets, while Table 20 includes additional methods and non-harmonized datasets. Table 19 demonstrates that the Harmonized-Data CDSSL method consistently outperforms both the Original CDSSL and the Random Initializer methods. In the Harmonized-ADNI dataset, Harmonized-Data CDSSL achieves an R^2 value of 0.63, compared to 0.40 for Original CDSSL and 0.05 for the Random Initializer. Table 20 introduces two additional methods: Augmentation CDSSL and Auxiliary CDSSL. These methods show strong performance, particularly on the ADNI dataset, where Augmentation CDSSL achieves an R^2 of 0.63 and Auxiliary CDSSL 0.62, both significantly outperforming Original CDSSL and Random Initializer.

The combined plot in Figure 14, visually compares these pretraining methods across original ADNI and Harmonized-ADNI. It highlights the consistent superiority of the Harmonized-Data CDSSL and Augmentation CDSSL methods, particularly on the ADNI and dataset A. This visual representation confirms that incorporating data harmonization and augmentation techniques significantly enhances model performance.

4.0 Discussion & Conclusion

This thesis presents a cross-domain self-supervised learning (CDSSL) framework for predicting the progression of Alzheimer's disease (AD) from MRIs, formulated as a regression task. Our pioneering effort aggregates a comprehensive set of internal and external cohorts to create a substantial dataset for model training. By using SimCLR pretraining on natural images followed by pretraining on medical images, we achieved the highest accuracy, effectively alleviating the domain shift challenge and greatly improving the generalization of the pretrained features. Our extensive experiments demonstrate the effectiveness of our approach in combating the lack of large-scale annotated data for training deep models for progression prediction. The best-performing model exhibits a substantial improvement over fully supervised models, demonstrating that the appropriate utilization of unlabeled images, including both natural and medical images, provides additional useful information that the model successfully learns from. Moreover, the proposed CDSSL approach can learn domain-invariant features, enhancing model generalization ability and robustness. This has the potential to identify patients at higher risk of progressing to AD and help develop better therapies at a lower cost to society. In our exploration of solutions for the effect of scanner variance, the significance of data harmonization in multi-center studies became evident. Properly harmonized data enhances the quality and reliability of research outcomes by ensuring consistency across varied equipment and protocols. We noted the benefits of data harmonization, from improving reproducibility to enhancing the statistical power of studies. However, challenges such as the potential of over-adjusting data or introducing biases underscore the importance of careful and validated approaches. Our study also delved into the potential of using scanner-specific attributes like the manufacturer and model in modeling. The performance of methods such as CDSSL, when compared against traditional models, showcased their capability in environments with diverse scanners. Using the Likelihood Ratio Test (LRT) provided deeper insights into model performance, facilitating better-informed decisions. Limitations

The datasets used in this study come from various sources with different criteria for

amyloid positivity and other clinical measures. This heterogeneity may introduce biases and affect the generalizability of the model across different populations and clinical settings.

Due to resource constraints, the analysis was limited to a subset of five slices per MRI scan. This limitation may hinder the model's ability to learn from the complete set of brain regions, potentially affecting its predictive performance and clinical relevance.

While the inclusion of scanner-specific information improved model performance, the study did not fully explore the impact of other scanner-related variables, such as imaging protocols, patient positioning, and software versions. These factors could further influence the results and model generalizability.

The study focuses on predicting AD progression over a 12-month period. This relatively short timeframe may not capture the full spectrum of disease progression, particularly for slower-progressing cases.

While the model shows promising results on various datasets, extensive clinical validation in real-world settings is still needed to ensure its practical applicability and reliability. Despite efforts to visualize model attention using GradCAM, the interpretability of deep learning models remains a challenge, potentially limiting their acceptance in clinical practice. Future Works

Future research should aim to include a more comprehensive set of MRI slices and additional imaging modalities (e.g., PET, DTI) to capture a broader range of brain regions and pathological features. This could potentially improve the model's predictive power and clinical relevance.

Further exploration of advanced harmonization techniques, including those that account for more scanner-related variables and imaging protocols, could improve the robustness and accuracy of the models across diverse clinical settings.

Incorporating longitudinal data over extended periods (e.g., 3-5 years) could provide deeper insights into the progression of AD and improve the model's ability to predict long-term outcomes, capturing both fast and slow-progressing cases.

Combining MRI data with other biomarkers, such as genetic information, cerebrospinal fluid analysis, and cognitive test scores, could enhance the predictive power and clinical relevance of the models. This multi-modal approach may provide a more comprehensive

view of AD progression.

Conducting prospective studies in diverse clinical settings will be crucial to validate the model's performance, assess its practical utility, and ensure its generalizability across different patient populations and healthcare systems.

Developing more advanced techniques for model interpretation and visualization could enhance the transparency and trustworthiness of the AI models, potentially increasing their adoption in clinical practice.

Exploring the applicability of the developed CDSSL techniques to other neurodegenerative diseases, such as Parkinson's disease or frontotemporal dementia, could expand the impact of this research. Automated Preprocessing Pipeline: Developing an end-to-end automated preprocessing pipeline that includes robust skull stripping, registration, and harmonization could streamline the clinical application of these models and reduce potential sources of variability.

Investigating federated learning approaches could allow for model training across multiple institutions without the need for data sharing, addressing privacy concerns and potentially increasing the diversity and size of the training dataset.

Developing models that can dynamically update predictions as new patient data becomes available could provide more personalized and timely prognostic information, enhancing clinical decision-making. By addressing these limitations and exploring these future directions, the potential of CDSSL and data harmonization in medical imaging can be further realized, leading to more accurate, reliable, and clinically relevant prognostic tools for Alzheimer's disease and potentially other neurological conditions.

Bibliography

Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., and R. Oleiwi, S. (2020). Towards a better understanding of transfer learning for medical imaging: a case study. *Applied Sciences*, 10(13):4523.

Ashburner, J. and Friston, K. J. (2005). Unified segmentation. *Neuroimage*, 26(3):839–851.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., MacWilliams, P., Mahdavi, S. S., Wulczyn, E., et al. (2022). Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488.

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S. E., Guo, Y., Matthews, P. M., and Rueckert, D. (2019). Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 541–549. Springer.

Bansal, R., Hao, X., and Peterson, B. S. (2017). Segmenting and validating brain tissue definitions in the presence of varying tissue contrast. *Magnetic resonance imaging*, 35:98–116.

Bardes, A., Ponce, J., and LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., et al. (2020). Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. *arXiv preprint arXiv:2010.05355*.

Bayer, J. M., Thompson, P. M., Ching, C. R., Liu, M., Chen, A., Panzenhagen, A. C., Jahanshad, N., Marquand, A., Schmaal, L., and Sämann, P. G. (2022). Site effects how-to and when: An overview of retrospective techniques to accommodate site effects in multisite neuroimaging analyses. *Frontiers in Neurology*, 13:923988.

Beer, J. C., Tustison, N. J., Cook, P. A., Davatzikos, C., Sheline, Y. I., Shinohara, R. T., Linn, K. A., Initiative, A. D. N., et al. (2020). Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, 220:117129.

Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., and Iglesias, J. E. (2021). Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution. *arXiv preprint arXiv:2107.09559*.

Blumberg, S. B., Palombo, M., Khoo, C. S., Tax, C. M., Tanno, R., and Alexander, D. C. (2019). Multi-stage prediction networks for data harmonization. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 411–419. Springer.

Blumberg, S. B., Tanno, R., Kokkinos, I., and Alexander, D. C. (2018). Deeper image quality transfer: Training low-memory neural networks for 3d images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 118–125. Springer.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924.

Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558.

Chang, X., Cai, X., Dan, Y., Song, Y., Lu, Q., Yang, G., and Nie, S. (2022). Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. *Physics in Medicine & Biology*, 67(14):145004.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020b). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Chen, X., Fan, H., Girshick, R., and He, K. (2020c). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, Z., Pawar, K., Ekanayake, M., Pain, C., Zhong, S., and Egan, G. F. (2023). Deep learning for image enhancement and correction in magnetic resonance imaging—state-of-the-art and challenges. *Journal of Digital Imaging*, 36(1):204–230.

Dadsetan, S., Hejrati, M., Wu, S., and Hashemifar, S. (2022). Robust alzheimer's progression modeling using cross-domain self-supervised deep learning. *arXiv preprint arXiv:2211.08559*.

Dagley, A., LaPoint, M., Huijbers, W., Hedden, T., McLaren, D. G., Chatwal, J. P., Papp, K. V., Amariglio, R. E., Blacker, D., Rentz, D. M., et al. (2017). Harvard aging brain study: dataset and accessibility. *Neuroimage*, 144:255–258.

Dar, S. U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., and Cukur, T. (2019). Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388.

Dhinagar, N. J., Thomopoulos, S. I., Rajagopalan, P., Stripelis, D., Ambite, J. L., Ver Steeg, G., and Thompson, P. M. (2023). Evaluation of transfer learning methods for detecting alzheimer's disease with brain mri. In *18th International Symposium on Medical Information Processing and Analysis*, volume 12567, pages 504–513. SPIE.

Dippel, J., Vogler, S., and Höhne, J. (2021). Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*.

Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430.

El-Sappagh, S., Abuhmed, T., Islam, S. R., and Kwak, K. S. (2020). Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. *Neurocomputing*, 412:197–215.

Ellis, J., Nathan, P. J., Villemagne, V. L., Mulligan, R., Saunder, T., Young, K., Smith, C. L., Welch, J., Woodward, M., Wesnes, K. A., et al. (2009). Galantamine-induced improvements in cognitive function are not related to alterations in $\alpha 4\beta 2$ nicotinic receptors in early alzheimer's disease as measured in vivo by 2-[18f] fluoro-a-85380 pet. *Psychopharmacology*, 202(1):79–91.

Fatania, K., Clark, A., Frood, R., Scarsbrook, A., Al-Qaisieh, B., Currie, S., and Nix, M. (2022). Harmonisation of scanner-dependent contrast variations in magnetic resonance imaging for radiation oncology, using style-blind auto-encoders. *Physics and Imaging in Radiation Oncology*, 22:115–122.

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

Fortin, J.-P., Sweeney, E. M., Muschelli, J., Crainiceanu, C. M., Shinohara, R. T., Initiative, A. D. N., et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132:198–212.

Franzmeier, N., Koutsouleris, N., Benzinger, T., Goate, A., Karch, C. M., Fagan, A. M., McDade, E., Duering, M., Dichgans, M., Levin, J., et al. (2020). Predicting sporadic alzheimer's disease progression via inherited alzheimer's disease-informed machine-learning. *Alzheimer's & Dementia*, 16(3):501–511.

Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., Van Der Walt, S., Descoteaux, M., Nimmo-Smith, I., and Contributors, D. (2014). Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

Graziani, M., Andrearczyk, V., and Müller, H. (2019). Visualizing and interpreting feature reuse of pretrained cnns for histopathology. In *Irish Machine Vision and Image Processing Conference (IMVIP 2019), Dublin, Ireland*.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Gu, R., Wang, G., Lu, J., Zhang, J., Lei, W., Chen, Y., Liao, W., Zhang, S., Li, K., Metaxas, D. N., et al. (2023). Cddsa: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation. *Medical Image Analysis*, 89:102904.

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

Guan, H., Liu, Y., Yang, E., Yap, P.-T., Shen, D., and Liu, M. (2021). Multi-site mri harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical image analysis*, 71:102076.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020a). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., and Xie, P. (2020b). Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv*.

Heker, M. and Greenspan, H. (2020). Joint liver lesion segmentation and classification via transfer learning. *arXiv preprint arXiv:2004.12352*.

Henaff, O. (2020). Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR.

Hosseinzadeh Taher, M. R., Haghighi, F., Feng, R., Gotway, M. B., and Liang, J. (2021). A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13. Springer.

Hu, D., Zeng, L.-L., Hu, D., and Zeng, L.-L. (2019). Multi-task learning of structural mri for multi-site classification. *Pattern Analysis of the Human Connectome*, pages 205–226.

Islam, K. T., Zhong, S., Zakavi, P., Chen, Z., Kavnoudias, H., Farquharson, S., Durbridge, G., Barth, M., McMahon, K. L., Parizel, P. M., et al. (2023). Improving portable low-field mri image quality through image-to-image translation using paired low-and high-field images. *Scientific Reports*, 13(1):21183.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., and Wang, Z. (2021). Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841.

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Kan, X., Dai, W., Cui, H., Zhang, Z., Guo, Y., and Yang, C. (2022). Brain network transformer. *Advances in Neural Information Processing Systems*, 35:25586–25599.

Karayumak, S. C., Bouix, S., Ning, L., James, A., Crow, T., Shenton, M., Kubicki, M., and Rathi, Y. (2019). Retrospective harmonization of multi-site diffusion mri data acquired with different acquisition parameters. *Neuroimage*, 184:180–200.

Keenan, K. E., Ainslie, M., Barker, A. J., Boss, M. A., Cecil, K. M., Charles, C., Chenevert, T. L., Clarke, L., Evelhoch, J. L., Finn, P., et al. (2018). Quantitative magnetic resonance imaging phantoms: a review and the need for a system phantom. *Magnetic resonance in medicine*, 79(1):48–61.

Kieselmann, J. P., Fuller, C. D., Gurney-Champion, O. J., and Oelfke, U. (2021). Cross-modality deep learning: contouring of mri data from annotated ct data only. *Medical physics*, 48(4):1673–1684.

Kurokawa, R., Kamiya, K., Koike, S., Nakaya, M., Uematsu, A., Tanaka, S. C., Kamagata, K., Okada, N., Morita, K., Kasai, K., et al. (2021). Cross-scanner reproducibility and

harmonization of a diffusion mri structural brain network: A traveling subject study of multi-b acquisition. *NeuroImage*, 245:118675.

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

Langhough Koscik, R., Hermann, B. P., Allison, S., Clark, L. R., Jonaitis, E. M., Mueller, K. D., Betthauser, T. J., Christian, B. T., Du, L., Okonkwo, O., et al. (2021). Validity evidence for the research category,"cognitively unimpaired–declining," as a risk marker for mild cognitive impairment and alzheimer's disease. *Frontiers in Aging Neuroscience*, 13:688478.

Lebedev, A. V., Westman, E., Beyer, M., Kramberger, M., Aguilar, C., Pirtosek, Z., and Aarsland, D. (2013). Multivariate classification of patients with alzheimer's and dementia with lewy bodies using high-dimensional cortical thickness measurements: an mri surface-based morphometric study. *Journal of neurology*, 260:1104–1115.

Li, H., Xue, F.-F., Chaitanya, K., Luo, S., Ezhov, I., Wiestler, B., Zhang, J., and Menze, B. (2021). Imbalance-aware self-supervised learning for 3d radiomic representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer.

Li, J., Zhou, P., Xiong, C., and Hoi, S. C. (2020). Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., and Yu, Y. (2019). Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10632–10641.

Liu, L., Drouet, V., Wu, J. W., Witter, M. P., Small, S. A., Clelland, C., and Duff, K. (2012). Trans-synaptic spread of tau pathology in vivo. *PloS one*, 7(2):e31302.

Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., and Jahanshad, N. (2021). Style transfer using generative adversarial networks for multi-site mri harmonization. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 313–322. Springer.

Liu, Q., Yu, L., Luo, L., Dou, Q., and Heng, P. A. (2020a). Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE transactions on medical imaging*, 39(11):3429–3440.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. (2020b). A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908.

Ma, Q., Zhang, T., Zanetti, M. V., Shen, H., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Fan, Y., Hu, D., Busatto, G. F., et al. (2018). Classification of multi-site mr images in the presence of heterogeneity using multi-task learning. *NeuroImage: Clinical*, 19:476–486.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 297–300. IEEE.

Misra, I. and Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717.

Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., Yeo, B. T., Initiative, A. D. N., et al. (2020). Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer.

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019). Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):1–16.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Petersen, R. C., Aisen, P., Beckett, L. A., Donohue, M., Gamst, A., Harvey, D. J., Jack, C., Jagust, W., Shaw, L., Toga, A., et al. (2010). Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368.

Qiu, A., Lee, A., Tan, M., and Chung, M. K. (2015). Manifold learning on brain functional networks in aging. *Medical image analysis*, 20(1):52–60.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32.

Reinhold, J. C., Dewey, B. E., Carass, A., and Prince, J. L. (2019). Evaluating the impact of intensity normalization on mr image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 890–898. SPIE.

Risacher, S. L., Saykin, A. J., Wes, J. D., Shen, L., Firpi, H. A., and McDonald, B. C. (2009). Baseline mri predictors of conversion from mci to probable ad in the adni cohort. *Current Alzheimer Research*, 6(4):347–361.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rueb, U., Stratmann, K., Heinsen, H., Seidel, K., Bouzrou, M., and Korf, H.-W. (2017). Alzheimer's disease: characterization of the brain sites of the initial tau cytoskeletal pathology will improve the success of novel immunological anti-tau treatment approaches. *Journal of Alzheimer's Disease*, 57(3):683–696.

Safari, M., Yang, X., and Fatemi, A. (2024). Mri data consistency guided conditional diffusion probabilistic model for mr imaging acceleration. In *Medical Imaging 2024: Clinical and Biomedical Imaging*, volume 12930, pages 202–205. SPIE.

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., Risacher, S. L., Nho, K., Huentelman, M. J., Craig, D. W., et al. (2010). Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimer's & Dementia*, 6(3):265–273.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., Andriole, K. P., and Michalski, M. (2018). Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer.

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

Singh, M., Sharma, S., Verma, A., and Sharma, N. (2017). Enhancement and intensity inhomogeneity correction of diffusion-weighted mr images of neonatal and infantile brain using dynamic stochastic resonance. *Journal of Medical and Biological Engineering*, 37:508–518.

Siqueira Pinto, M., Winzeck, S., Kornaropoulos, E. N., Richter, S., Paolella, R., Correia, M. M., Glocker, B., Williams, G., Vik, A., Posti, J. P., et al. (2023). Use of support vector machines approach via combat harmonized diffusion tensor imaging for the diagnosis and prognosis of mild traumatic brain injury: a center-tbi study. *Journal of Neurotrauma*.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97.

Sowrirajan, H., Yang, J., Ng, A. Y., and Rajpurkar, P. (2021). Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR.

Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., and Dickscheid, T. (2018). Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 663–671. Springer.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Stonnington, C. M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack Jr, C. R., Chen, K., Ashburner, J., and Frackowiak, R. S. (2008). Interpreting scan data acquired from multiple scanners: a study with alzheimer's disease. *Neuroimage*, 39(3):1180–1185.

Sun, Y., Yuan, P., and Sun, Y. (2020). Mm-gan: 3d mri data augmentation for medical image segmentation via generative adversarial networks. In *2020 IEEE International conference on knowledge graph (ICKG)*, pages 227–234. IEEE.

Svanera, M., Savardi, M., Signoroni, A., Benini, S., and Muckli, L. (2024). Fighting the scanner effect in brain mri segmentation with a progressive level-of-detail network trained on multi-site data. *Medical Image Analysis*, 93:103090.

Tian, X., Liu, J., Kuang, H., Sheng, Y., Wang, J., and Initiative, T. A. D. N. (2022). Mri-based multi-task decoupling learning for alzheimer's disease detection and mmse score prediction: A multi-site validation. *arXiv preprint arXiv:2204.01708*.

Tian, Y., Krishnan, D., and Isola, P. (2020a). Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020b). What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839.

Timmermans, C., Smeets, D., Verheyden, J., Terzopoulos, V., Anania, V., Parizel, P. M., and Maas, A. (2019). Potential of a statistical approach for the standardization of multi-center diffusion tensor data: a phantom study. *Journal of Magnetic Resonance Imaging*, 49(4):955–965.

Tomar, D., Bozorgtabar, B., Lortkipanidze, M., Vray, G., Rad, M. S., and Thiran, J.-P. (2022). Self-supervised generative style transfer for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1998–2008.

Tong, Q., Gong, T., He, H., Wang, Z., Yu, W., Zhang, J., Zhai, L., Cui, H., Meng, X., Tax, C. W., et al. (2020). A deep learning–based method for improving reliability of multicenter diffusion kurtosis imaging with varied acquisition protocols. *Magnetic Resonance Imaging*, 73:31–44.

Torbati, M. E., Minhas, D. S., Ahmad, G., O'Connor, E. E., Muschelli, J., Laymon, C. M., Yang, Z., Cohen, A. D., Aizenstein, H. J., Klunk, W. E., et al. (2021). A multi-scanner neuroimaging data harmonization using ravel and combat. *Neuroimage*, 245:118703.

Tournier, J.-D., Calamante, F., Gadian, D. G., and Connelly, A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted mri data using spherical deconvolution. *Neuroimage*, 23(3):1176–1185.

Tudorascu, D. L., Karim, H. T., Maronge, J. M., Alhilali, L., Fakhran, S., Aizenstein, H. J., Muschelli, J., and Crainiceanu, C. M. (2016). Reproducibility and bias in healthy brain segmentation: comparison of two popular neuroimaging platforms. *Frontiers in neuroscience*, 10:503.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320.

Vendrow, E. and Schonfeld, E. (2022). Understanding transfer learning for chest radiograph clinical report generation with modified transformer architectures. *arXiv preprint arXiv:2205.02841*.

Venugopalan, J., Tong, L., Hassanzadeh, H. R., and Wang, M. D. (2021). Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):1–13.

Vovk, U., Pernus, F., and Likar, B. (2007). A review of methods for correction of intensity inhomogeneity in mri. *IEEE transactions on medical imaging*, 26(3):405–421.

Wang, X., Chen, H., Ran, A.-R., Luo, L., Chan, P. P., Tham, C. C., Chang, R. T., Mannil, S. S., Cheung, C. Y., and Heng, P.-A. (2020). Towards multi-center glaucoma oct image screening with semi-supervised joint structure and function multi-task learning. *Medical Image Analysis*, 63:101695.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Xie, H., Shan, H., Cong, W., Zhang, X., Liu, S., Ning, R., and Wang, G. (2019). Dual network architecture for few-view ct-trained on imagenet data and transferred for medical imaging. In *Developments in X-ray Tomography XII*, volume 11113, pages 184–194. SPIE.

Xu, L., Pearlson, G., and Calhoun, V. D. (2009). Joint source based morphometry identifies linked gray and white matter group differences. *Neuroimage*, 44(3):777–789.

Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. (2019). Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR.

Zhang, C., Moeller, S., Demirel, O. B., Uğurbil, K., and Akçakaya, M. (2022). Residual raki: A hybrid linear and non-linear approach for scan-specific k-space deep learning. *NeuroImage*, 256:119248.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

Zhang, Y.-D., Dong, Z., Wang, S.-H., Yu, X., Yao, X., Zhou, Q., Hu, H., Li, M., Jiménez-Mesa, C., Ramirez, J., et al. (2020). Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 64:149–187.

Zhao, F., Wu, Z., Zhu, D., Liu, T., Gilmore, J., Lin, W., Wang, L., and Li, G. (2023). Disentangling site effects with cycle-consistent adversarial autoencoder for multi-site cortical data harmonization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 369–379. Springer.

Zhong, J., Wang, Y., Li, J., Xue, X., Liu, S., Wang, M., Gao, X., Wang, Q., Yang, J., and Li, X. (2020). Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: application to neonatal white matter development. *Biomedical engineering online*, 19(1):1–18.

Zhou, H.-Y., Yu, S., Bian, C., Hu, Y., Ma, K., and Zheng, Y. (2020). Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–407. Springer.

Zhou, Z., Sodha, V., Pang, J., Gotway, M. B., and Liang, J. (2021). Models genesis. *Medical image analysis*, 67:101840.

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., and Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492.

Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S. K., and Zheng, Y. (2020). Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical image analysis*, 64:101746.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., and Zheng, Y. (2019). Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–428. Springer.

Zuo, L., Dewey, B. E., Carass, A., Liu, Y., He, Y., Calabresi, P. A., and Prince, J. L. (2021a). Information-based disentangled representation learning for unsupervised mr harmonization. In *International Conference on Information Processing in Medical Imaging*, pages 346–359. Springer.

Zuo, L., Dewey, B. E., Liu, Y., He, Y., Newsome, S. D., Mowry, E. M., Resnick, S. M., Prince, J. L., and Carass, A. (2021b). Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569.