Comparative Analysis of Generalization and Harmonization Methods for 3D Brain fMRI Images: A Case Study on OpenBHB Dataset

Soroosh Safari Loaliyan
University of California Riverside
University Avenue, Riverside, CA
ssafa013@ucr.edu

Greg Ver Steeg
University of California Riverside
University Avenue, Riverside, CA
gregory.versteeg@ucr.edu

Abstract

Due to limitations on the amount of available data from a single source, combining data from different sources can significantly improve the statistical analysis of fMRI images. However, because the training and target sources are usually different, applying a deep learning model trained on the source domain leads to inconsistent results when applied to the target domain, especially on 3D MRI images due to variations in bias across scanners. While Harmonization methods and Domain Adaptation (DA) are popular approaches for handling multi-site MRI data, Domain Generalization (DG) is less studied and may offer some unique benefits.

In this study, we explore the impact of DG, and compare data harmonization on brain age prediction using 3D fMRI images from the OpenBHB dataset. The dataset consists of 3D T1 brain MRI scans aggregated from 10 *publicly available datasets, comprising* $N = 3985$ *individuals and acquired on more than* 70 *different scanners. We focus on the Big Healthy Brains (BHB) dataset, utilizing Voxel-Based Morphometry (VBM) images and performing a split into training, out-of-domain validation, and out-of-domain test sets based on site metadata. Our experiments involve training models based on the architecture introduced in [8], extending it as per [9], and evaluating various generalization techniques on both harmonized and non-harmonized data and their differences. Our findings highlight that harmonization increases the Mean Absolute Error (MAE) for all methods. DA and DG methods reliably improve performance on out-of-domain tests, with Domain Adversarial Neural Network (DANN) exhibiting the best performance. This paper reports the complex interplay between data harmonization and model generalization, providing insights into the selection and application of generalization techniques in neuroimaging.*

1. Introduction

In the field of neuroscience machine learning, models need to be both effective and reliable, meaning they should be sensitive to biological factors while remaining unaffected by non-biological variables such as site or instrument variations. Magnetic Resonance Image (MRI) is a classic tool with many applications [3]. However, the process of acquiring these scans and the difference between devices can affect the appearance of healthy and abnormal tissues. Convolutional Neural Networks (CNNs) have outstanding results on many medical imaging tasks but they are sensitive to the differences in Imaging protocols. In other words, a CNN model trained on a particular imaging protocol is less effective on the images from other imaging sources because of domain shift [12]. In neuroimaging research, addressing the challenges posed by domain (scanner) shifts in 3D fMRI brain images is important for the development of robust and reliable models.

To address these challenges, many different solutions have been introduced including Domain Adaptation (SDA) which needs both the labeled source domain and labeled target domain. Practically this method is not effective because it is not possible to have all of the required information from different sites and devices, thus accessing the labeled information from the target domain is not a realistic approach. Another method is Unsupervised Domain Adaptation (UDA) where the access to the target domain is limited to unlabeled data. This method is more practical than SDA but it is also not applicable to the Medical Imaging Analysis field (MedIA) because the access to target domain is often impossible. For example, where we want to use prediction model on a new installed device from a new site. Finally, the last method is Domain Generalization, in this scenario, there is no access to the target domain data (labeled or unlabeled). This method uses different ways to predict the effect of target domain constrains using multiple source domain data. This method is more practical in many ways. First, it can generalize the model to any unseen target domain.

Thus, the final model can be used to predict the required information on any data from unseen domains. Additionally, harmonization techniques primarily involve preprocessing input image data from various domains to produce altered images that appear more alike, yet still retain the important attributes/information necessary for distinguishing prediction outcomes. The issue with harmonization methods is that they must be reapplied to new images from any source, and this process can alter key image attributes in a way that potentially increases prediction error.

In this context, domain shifts refer to differences in scanners and image protocols across different sites. Some examples of domain shift include flip angle, acquisition orientation, and slice thickness. Therefore, MRI may differ in many aspects from center to center or study to study. Dealing with these parameters is impossible in many scenarios where we don't have the details of each scanner and protocol.

This study aims to thoroughly investigate the applicability and performance of well-known domain adaptation and generalization methods on both harmonized and non-harmonized data. By comparing these methods against each other and against the current state-of-the-art approach, we seek to provide insights into their efficacy in handling domain shifts in heterogeneous datasets.

2. Background

Many Studies have shown the impact of different techniques on the prediction of target values (such as age, cancer, etc.) in different medical domains. H.Guan, and M.Liu [6], summarized many domain adaptation techniques in Medical Image Analysis (MedIA). As mentioned by them, there is not too much task-specific research on this topic. Although many of these researches try to show the improvement of their adaptation method based on the source and target domain data on 2D images, there is few studies that work with 3D fMRI images. Another problem mentioned by them is the "Multi-Source/Multi-Target Domain Adaptation" problem. Current domain adaptation (DA) methods typically concentrate on adapting from a single source domain, meaning they train a model using data from one domain. However, in practical scenarios, there could be several source domains (such as various imaging centers).

Many other research studies work on DG methods. As mentioned by [21], these studies primarily use data from the source domain to adapt the model, enhancing its performance on unseen domains without requiring any data from those domains. These methods are more practical on real neuroimaging tasks because while they don't use any data from the target domain, these methods mainly use multiple source domain data.

There are many different approaches on UDA like discrepancy-based approaches such as Deep Adaptation Network (DAN) and Correlation Alignment (CORAL) [18], adversarial-based approaches such as Domain Adversarial Neural Network (DANN) [17] and Adversarial Discriminative Domain Adaptation (ADDA) [19], moment matching approaches like Moment Matching for Multi-Source Domain Adaptation (M3SDA) [15], Domain Genrelaization (DG) such as Domain Adversarial Neural Network with Cooperative Examples (DANNCE) [17] and Harmonization techniques such as Combat [4, 10].

Many researches state that the challenges in domain generalization stem from various factors including the difficulty of achieving domain invariance across unseen domains, which can lead to increased prediction errors when the invariance achieved on training domains does not generalize well. Furthermore, assumptions made by algorithms about domain invariance and the relevance of chosen datasets might not accurately reflect real-world scenarios, complicating the evaluation of these methods. The lack of a standardized framework for model selection and evaluation poses additional hurdles, as it becomes challenging to fairly compare different methods or to choose the best hyperparameters without a suitable validation set. Additionally, the strong performance of Empirical Risk Minimization (ERM) compared to more complex domain generalization strategies, when combined with modern architectures and careful tuning, questions the necessity of domain-specific approaches over enhancements to in-distribution generalization techniques [5, 7, 13, 14].

3. Methodology

The core of our research lies in the exploration of generalization methods from the Stanford WILDS package [11, 16]. It includes Empirical Risk Minimization (ERM), Invariant Risk Minimization (IRM) [2], Domain Adversarial Neural Network (DANN) [17], Deep Correlation Alignment (deepCORAL) [18]. Additionally we evaluated Domain Adversarial Neural Network with Cooperative Examples (DANNCE) which is a generalization method presented by A.Sicila [17]. We have implemented all of the above methods on harmonized and non-harmonized data, to predict brain age from fMRI images. For this task, we use the ComBat harmonization technique which is a well-known fMRI image harmonization in MedIA [4, 10]. We adopt a model architecture introduced in [8], further extended as per [9], to capture spatial information from 3D fMRI images. Our objective is to assess the effectiveness of these methods in handling domain shifts and improving the generalizability of brain age prediction models across diverse datasets.

Our study conducts a comparative analysis of generalization methods on the OpenBHB dataset, with an emphasis on the harmonization process's influence. Previous studies show that harmonization methods normally decrease accuracy or increase the error. As mentioned by some researches [1, 20], while harmonization techniques like ComBat and cVAE can effectively reduce dataset differences to pool MRI data from multiple sources, they may inadvertently also remove biologically relevant information. This loss of information could be detrimental to the models trained on the harmonized data, potentially leading to increased prediction errors in tasks such as predicting Mini-Mental State Examination (MMSE) scores and clinical diagnoses. Our results also indicate an overall increase in MAE, suggesting the potential loss of predictive variance. We use multi-source, multi-target domain dataset to measure the performance of known DG methods. we compare MAE from different generalization methods on prediction of brain age from fMRI images. The domain generalization methods that we use in this work are:

Empirical Risk Minimization (ERM): In the context of domain generalization, ERM involves combining data from multiple source domains and training a model to minimize this combined empirical risk, with the hope that it generalizes well to unseen domains. The primary focus of ERM is on achieving the lowest possible error on the training data, often without explicitly considering the variability or differences between domains. This approach is straightforward and widely used, but it can be limited by its reliance on the assumption that minimizing empirical risk on the source domains will ensure generalization to target domains, potentially overlooking the domain-specific nuances that might not generalize well.

Invariant Risk Minimization (IRM): The goal of IRM is to identify and leverage the underlying causal structures within the data that remain constant across domains, thus enabling the model to generalize better to unseen domains. This is achieved through an optimization process that not only minimizes the empirical risk but also imposes constraints to ensure the invariance of the model's predictions across domains. Unlike ERM, which primarily focuses on aggregate loss, IRM specifically targets the robustness and transferability of learned features by enforcing that the optimal predictor should perform consistently across all known environments. This focus on invariance to domain shifts makes IRM particularly suited for situations where domain variability is significant and where uncovering causal relationships is crucial for generalization.

Domain Adversarial Neural Network (DANN): The core idea behind DANN is to train a model in a way that makes it difficult to distinguish between source domain data (on which the model is trained) and target domain data (on which the model is tested), thereby minimizing domain discrepancy. This is achieved through a unique architecture that combines a feature extractor, a task-specific predictor, and a domain classifier in an adversarial training framework. The feature extractor learns to generate domain-invariant features, while the domain classifier tries to distinguish between the source and target domain features. Simultaneously, the task-specific predictor focuses on the primary learning task (e.g., classification or regression). Through backpropagation and gradient reversal layers, the model is encouraged to find feature representations that are both useful for the main task and indistinguishable by the domain classifier, thus promoting domain invariance and enhancing the model's ability to generalize from the source to the target domain.

Deep Correlation Alignment (deepCORAL): deepCORAL focuses on aligning the second-order statistics (i.e., covariance) of source and target domain feature distributions to reduce the domain shift. The essence of Deep CORAL lies in its ability to minimize the discrepancy between the source and target domains by adjusting the deep neural network's features such that the correlation matrices of the two domains are brought closer. This is accomplished through a loss function that quantifies the difference in covariance between features extracted from the source and target domains, encouraging the network to learn representations that are invariant across domains. By integrating this alignment objective with the standard task-specific loss (e.g., classification or regression loss), Deep CORAL effectively guides the network to not only perform well on the source domain task but also to generalize better to the target domain by mitigating the impact of domain-specific variations. This approach leverages the deep learning framework's capacity for feature extraction and representation learning, making it a powerful tool for domain adaptation challenges where the goal is to bridge the gap between differently distributed datasets without requiring explicit domain labels during training.

Domain Adversarial Neural Network with Cooperative Examples (DANNCE): DANNCE differs from DANN primarily in its approach to addressing domain generalization challenges. While DANN focuses on minimizing domain discrepancies through adversarial training to learn domain-invariant features for domain adaptation, DANNCE goes further by enhancing source diversity and tackling the broader challenges of domain generalization. Specifically, DANNCE implements a strategy to generate examples that target the weaknesses of the feature extractor, aiming to deceive the domain discriminator in a way that promotes cooperative adaptation. This results in a more diversified source representation, pushing the model towards learning features that are inherently more generalizable across various unseen domains. By actively manipulating source domain data to better align with the domain discriminator's goals, DANNCE not only seeks domain invariance like DANN but also enhances the model's adaptability and performance across different domains by enriching the representation of source domain features. This approach marks a significant strategic shift, focusing on the dynamic enrichment of source data to overcome domain generalization challenges more effectively.

Although DANN and deepCORAL are known as DA techniques, meaning they access the data from the target domain, in this work, we split the training dataset into two parts such that both labeled and unlabeled datasets for DANN and deepCORAL are from the training dataset. this means that in our work, we use DANN and deepCORAL as domain generalization techniques because none of the domains are shared between the training, validation, and test datasets. Also, the domain between labeled and unlabeled datasets is not shared.

Additionally, we investigate the effectiveness of ComBat harmonization, a powerful post-processing technique designed to remove technical between-scanner variation. ComBat successfully removes inter-site technical variability as demonstrated in recent papers [4, 10].

3.1. Dataset

We use the Big Healthy Brains (BHB) dataset, a comprehensive aggregation of 3D T1 brain MRI scans from healthy controls (HC). This dataset encompasses data from 70 different scanners, comprising $N = 3985$ individuals and unifying various preprocessing techniques, including VBM.

For our experiments, we exclusively use VBM images from the BHB dataset. To simulate domain shifts, we perform a split into three categories: training, out-of-domain validation, and out-of-domain test sets. This split is achieved by selecting sites, ensuring 70% of the data points for training, 15% for validation, and 15% for testing Tab. 1.

The training dataset is divided into two parts for better comparison between DA (DANN and deepCORAL) and others. As you can see in Tab. 1, we divided the training data into T1 as labeled for all generalization methods and T2 for methods that accept unlabeled data. T1 has 40% of all data while T2 has 30%. To make a fair comparison between all generalizations, we compare all methods in two situations, the first situation assumes that all of the methods are using the whole training dataset (T1 plus T2), while the second comparison uses only T1 as the labeled data for generalization methods that only accept labeled data (ERM, IRM, DANNCE) and for DANN and deepCORAL, we use T1 as labeled and T2 and unlabeled data. We do all of these comparisons on both harmonized and non-harmonized data.

Each dataset has different domains (sites). Table 1 shows T1 images are from 16 first sites while T2 includes 34 to 36, 46 to 64, and site 69. validation dataset sites are in the range of 15 to 33 and test sites are from 37 to 45 plus 65. we did this division to make sure that the total number of data points for each dataset is as close as possible to the division percentage previously mentioned.

Though DANN and deepCORAL are well-known DA techniques and these methods should have access to the unlabeled target domains, this site distinction makes sure that these methods are being used as generalization methods in this study. because there is no intersection between D_{train}, D_{test}, and D_{val} in other words:

$$D_{train} \cap D_{test} = D_{train} \cap D_{val} = D_{test} \cap D_{val} = \emptyset \qquad (1)$$

4. Results

Our comprehensive investigation into the effectiveness of generalization methods on the mixed VBM dataset, sourced from the extensive Big Healthy Brains (BHB) dataset, highlights significant findings. Leveraging the sophisticated architecture introduced by Gupta et al. [8], the study underscores the model's adeptness at navigating domain shifts and safeguarding spatial data integrity.

The model has been trained for 50 epochs, with a calibrated learning rate of 0.01 and a weight decay parameter of 0.001. Employing the Mean Squared Error (MSE) as the loss criterion and the Adam optimization algorithm, we evaluated model performance, employing the Mean Absolute Error (MAE) as our primary metric.

A evaluation approach was adopted to assess the comparative quality of DANN/deepCORAL against other domain generalization (DG) methodologies, demand the exploration of two distinct scenarios. In the first scenario, a comprehensive labeling of the T1 and T2 datasets ease a direct comparison across all DG methods, integrating DANN and deepCORAL within this framework (Fig. 1, Fig. 2). The subsequent scenario introduced a presentation between labeled (T1) and unlabeled (T2) datasets, a configuration that uniquely positioned DANN and deepCORAL to leverage unlabeled data, thus simulating a more complex domain adaptation challenge (Fig. 3, Fig. 4).

4.1. First Scenario (T1 plus T2 as labeled)

In this scenario, the entire dataset, comprising both T1 and T2, is treated as labeled, providing a comprehensive basis for evaluating the effectiveness of various DG methods under uniform labeling conditions. This setup offers a unique vantage point for assessing each method's capability in leveraging labeled data for domain adaptation and prediction.

Our first line of investigation under this scenario focused on non-harmonized data. This analysis was essential in understanding how each generalization method copes with data diversity without any harmonization efforts to

Table 1. Data and Site Distribution among Train, Validation, and Test dataset

Data Distribution	**Train Subset 1 (40%)**	**Train Subset 2 (30%)**	**Validation (15%)**	**OOD Test (15%)**	**Total**
Sites	0-14	34-36, 46-64, 69	15-33	37-45, 65	70
Data	1589	1207	594	594	3984

	val	train	test
IRM (Without T2)	10.489749	6.085962	9.575923
ERM (Without T2)	9.557683	3.154767	9.537376
DANNCE (Without T2)	9.797408	3.057051	**9.448741**
DANN (T2 as Unlabeled)	**9.437050**	**0.831619**	9.535244
deepCORAL (T2 as Unlabeled)	10.529750	5.890404	9.528206

Table 2. Harmonized Data Average MAE Over Last 20 epochs

minimize domain shifts. Remarkably, all DG methods demonstrated superior performance compared to the baseline method (ERM) across the testing dataset, indicating their inherent strength in extracting and utilizing domain-specific information from labeled data effectively.

Among the evaluated methods, DANN emerged as the standout performer, achieving the lowest MAE values across training, validation, and testing datasets (Fig. 1). This outcome not only highlights DANN's proficiency in domain adaptation using labeled data but also underscores its capacity to detect and grasp domain-specific distinctions more effectively than its counterparts. Contrarily, DANNCE, while not utilizing the full spectrum of the training dataset for domain prediction tasks, adeptly generates additional examples from subsets of the training data, enhancing its domain discrimination capabilities.

The exploration extends into the domain of harmonized data, where the effects of data uniformity on model performance are meticulously scrutinized (Fig. 2). Within this context, DANN maintains its position as the method with the lowest MAE for the training dataset. However, a notable shift occurs in the performance dynamics during validation and testing phases, where DANN exhibits increasing MAE values over epochs, signaling a tendency towards overfitting in a harmonized data environment.

In stark contrast, DANNCE demonstrates remarkable resilience against overfitting, quickly stabilizing its performance and outpacing other methods in adapting to harmonized data. This stability can be attributed to DANNCE's strategic use of generated examples for domain prediction, a technique that proves particularly effective in navigating the reduced inter-domain variability characteristic of harmonized datasets.

4.2. Second Scenario (T1 Labeled, T2 Unlabeled)

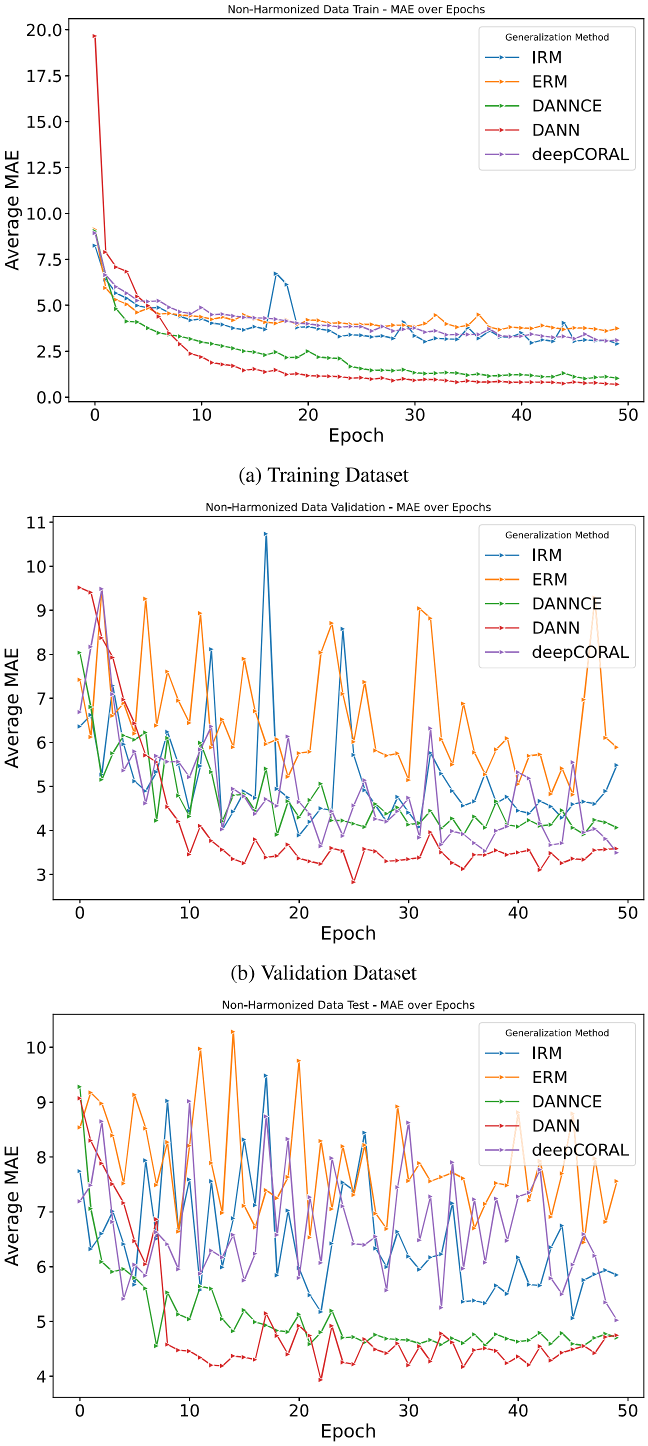
In this scenario, we divide the train dataset into two distinct parts: T1, comprising labeled data, and T2, containing unlabeled data. This division sets the stage for a different evaluation, particularly for DANN and deepCORAL. These methods, designed for domain adaptation, uniquely accept unlabeled data, presuming such data originates from the target domain—a notion not entirely applicable in our experimental setup.

Our initial experiments in this scenario explored the dynamics of non-harmonized data, meticulously capturing the performance of various generalization methods across training, validation, and testing phases. Specifically, DANN demonstrated remarkable efficiency, securing the lowest MAE values across all datasets, thereby underscoring its robustness in imposing available labeled data for domain prediction. Conversely, deepCORAL struggled, manifesting the highest MAE values, even lagging behind the established baseline approach (ERM). This outcome was notable, given deepCORAL's access to unlabeled data, suggesting potential limitations in its data utilization strategy. Meanwhile, DANNCE showcased comparable efficacy to DANN, albeit without relying on any unlabeled data, hinting at its adeptness in maximizing the utility of labeled data alone (Fig. 3).

For a more better comparison on harmonized data, we focused towards analyzing the average MAE over the concluding 20 epochs. The outcomes, detailed in Tab. 2, reveal a layered narrative of methodological performance on the harmonization process.

Within this framework, DANN consistently outperformed other methods in the realms of training and validation, showcasing its adeptness at navigating the difficulties of harmonized data without resulting to overfitting—a common pitfall observed in the previous scenario. This resilience is particularly noteworthy, given the identical volume of data leveraged in both harmonized and non-harmonized contexts, underscoring DANN's strategic employment of unlabeled data predominantly for domain discrimination purposes (Fig. 4).

Table 2 shows these findings, presenting a view of each method's average performance across the last 20 epochs on harmonized data. Here, DANN and DANNCE are as particularly robust, demonstrating minimal performance fluctuations and well stability across validation and testing phases—attributes indicative of their enhanced adaptability and the strategic foresight embedded within their operational paradigms.

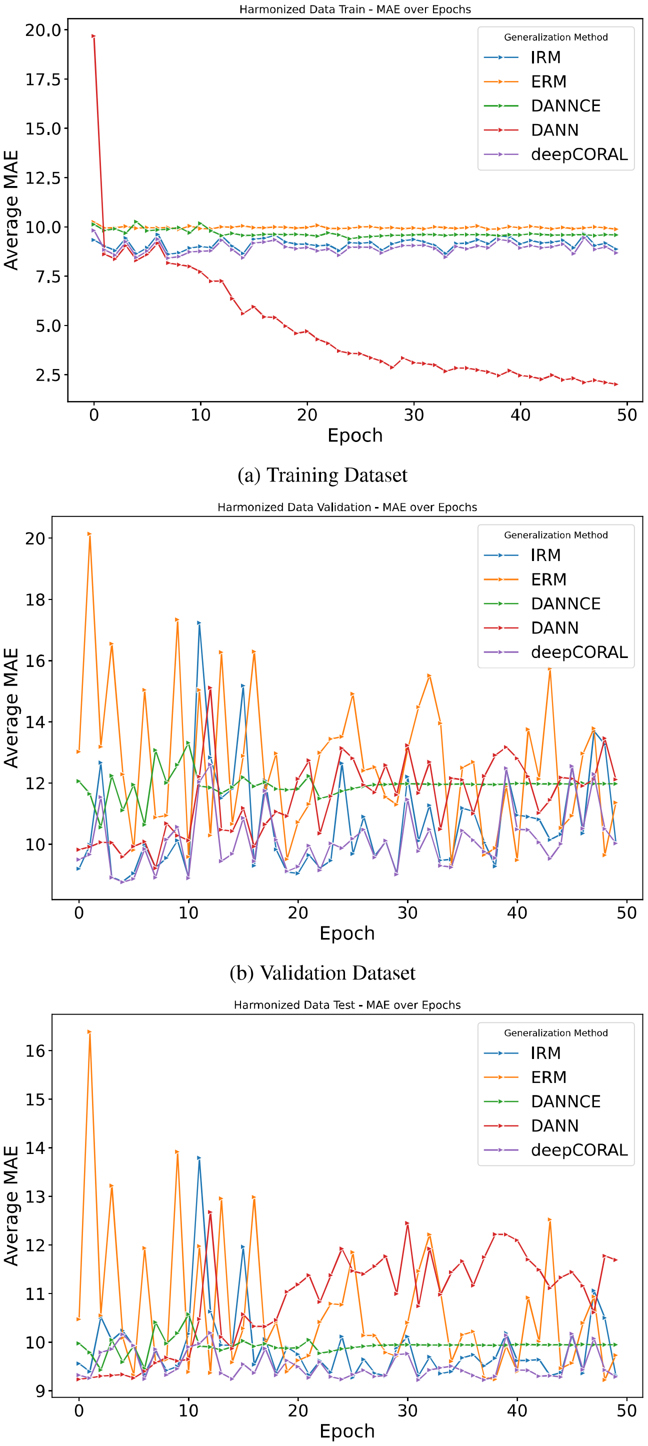


(a) Training Dataset

(b) Validation Dataset

(c) Test Dataset

Figure 1. MAE performance on Non-Harmonized Data



(a) Training Dataset

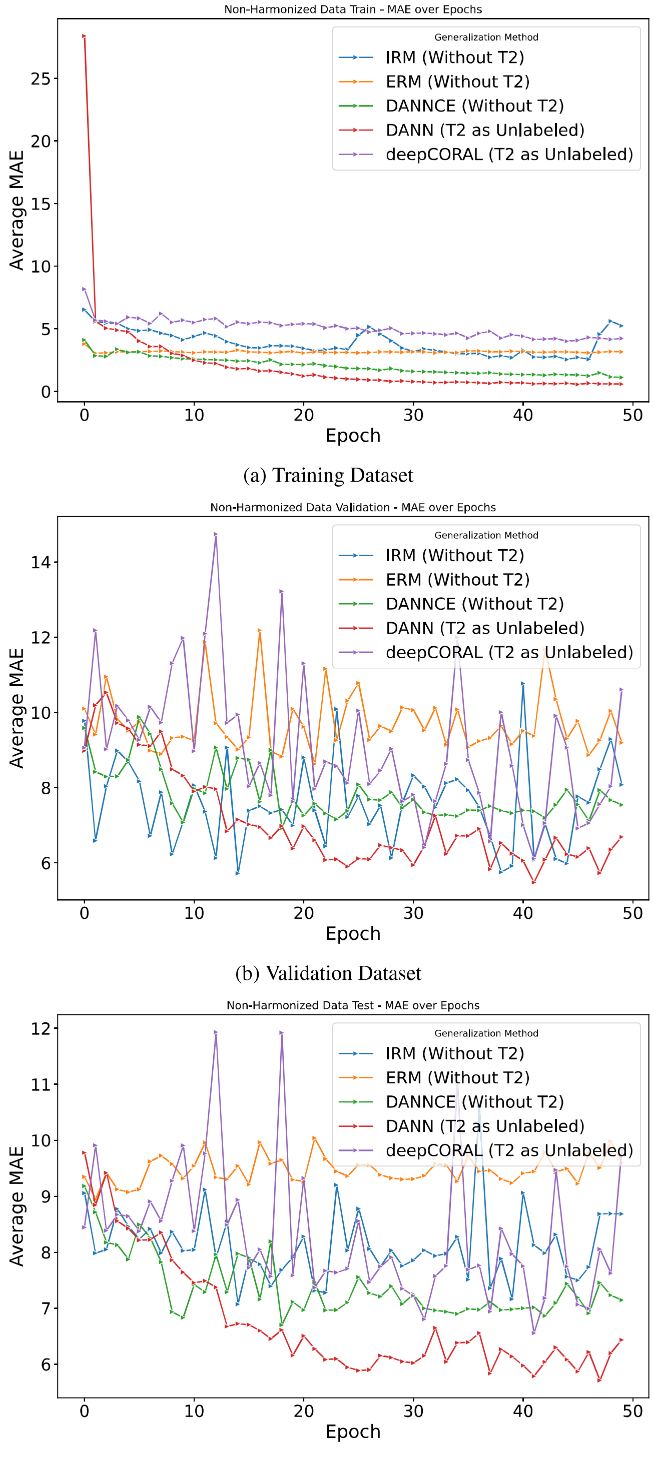
(b) Validation Dataset

(c) Test Dataset

Figure 2. MAE performance on Harmonized Data

4.3. General Compare Between Scenarios

Upon comparing the outcomes from both scenarios, several interesting findings emerge, showing the efficacy of DG

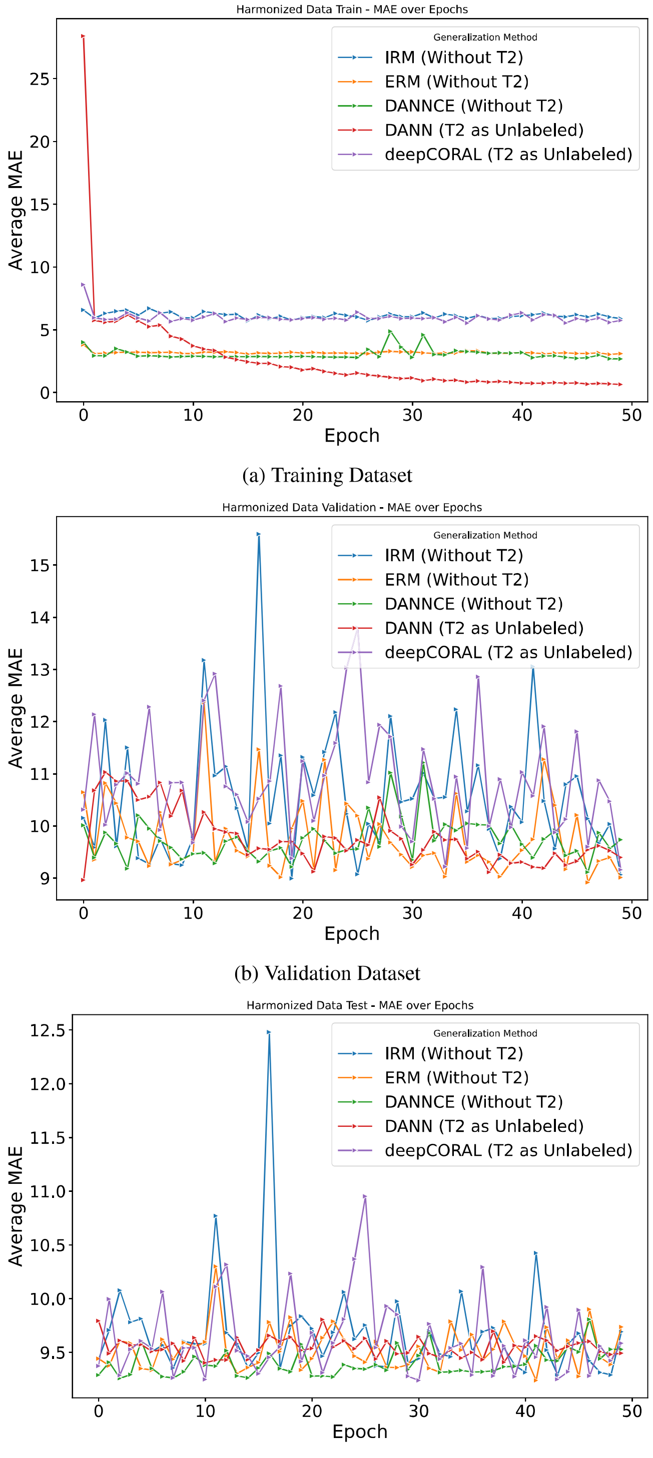


(a) Training Dataset

(b) Validation Dataset

(c) Test Dataset

Figure 3. MAE performance on Non-Harmonized Data (Second Scenario)



(a) Training Dataset

(b) Validation Dataset

(c) Test Dataset

Figure 4. MAE performance on Harmonized Data (Second Scenario)

methods in handling 3D MRI images under varied conditions of data labeling and harmonization. This comparison not only illuminates the inherent adaptability of these methods but also offers valuable insights into their operational dynamics across different data processing landscapes.

- **DANNCE's Performance with Less Data:** An observation from the analysis is DANNCE's improved performance in the harmonized test dataset within the second scenario, despite utilizing a lesser volume of training data. This finding challenges the conventional suggestions that an increased quantity of training data invariably leads to better model performance. DANNCE's ability to achieve lower MAE with fewer data points underscores the method's efficiency in extracting and leveraging critical domain-specific information, suggesting that the quality of data and strategic data usage can significantly influence model efficacy.
- **Impact of Harmonization on MAE:** The harmonization process, intended to minimize inter-domain variability, unintentionally leads to the obscuration of vital predictive information. This effect is evidenced by the increased MAE values post-harmonization, indicating a potential compromise in the predictive integrity of the models. Moreover, the harmonization reduces the relative impact of the training data volume on model performance. The lack of significant differences in the test dataset MAE across scenarios, despite a considerable disparity in training data volume, highlights the reduction on the returns of increased data volume in the context of harmonized datasets.
- **Data Volume and Non-Harmonized Data Results:** The quantity of training data exerts a pronounced influence on model performance with non-harmonized data. For instance, in the first scenario, where the training dataset is larger, DANNCE and DANN exhibit an average MAE of 4 to 5 on the test dataset. On the contrary, in the second scenario, with a reduced data volume, the average MAE increases to between 6 and 7. This variation distinctly illustrates the significance of data volume in enhancing model performance, particularly when dealing with non-harmonized data.
- **Comparative MAE of DANN and DANNCE:** Another insightful observation is the elimination of the difference in test dataset MAE between DANN and DANNCE as more labeled data is employed for both methods in non-harmonized conditions. This convergence suggests that the gap in performance between these methods narrows with the availability of more comprehensive labeled data, highlighting the potential for data volume to mitigate performance discrepancies between different DG methods.
- **Improvements in Prediction Tasks:** Both experimental scenarios demonstrate the generalization methods' effectiveness in improving prediction tasks on non-harmonized data. This effectiveness is particularly pronounced in the context of 3D MRI images, where DG methods show a remarkable capacity to adapt to and leverage the inherent complexity and variability of the data for enhanced predictive accuracy.
- **Stability of DANN and DANNCE:** The analysis further reveals that DANN and DANNCE exhibit less fluctuation in average MAE during the validation and testing processes. This stability observed even with fewer data points, suggests that these methods have inherent robustness and adaptability, enabling them to maintain consistent performance levels across different phases of model evaluation.

5. Conclusion

Our study compares different ways to make fMRI data work well for predicting images, whether the data is harmonized or not. We found that DANN is better than other methods because it can create features that work across different areas without getting confused by changes in the data. DANN and its version, DANNCE, use a special kind of training to do this, which is better than just matching data statistics or adapting to new data on the fly. These methods, especially DANN, are also good because they don't need new, unlabeled data from the target area, which is hard to get for fMRI studies. While harmonization needs to be redone for every new image and domain adaptation struggles without enough relevant data, DANN and DANNCE stand out for their ability to predict accurately across different domains. Our results suggest using DANN and DANNCE more in fMRI prediction is a promising path for future research, showing they offer clear advantages over traditional Domain Adaptation and Harmonization methods.

6. Future Work

Our research provides valuable insights into the efficacy of these methods, particularly in the context of brain age prediction, it also opens several avenues for future research.

Feature Attribution Analysis: Using advanced methods like SHAP and LIME to understand which image features impact model predictions can reveal site-specific patterns affecting data consistency. This helps in creating stronger methods for generalizing across different sites.

Deep Learning for Feature Discovery: Using unsupervised and semi-supervised models like autoencoders and GANs could help discover hidden features in MRI images, revealing complex patterns not seen with standard analysis.

Cross-Site Variability Analysis: By doing a systematic variability analysis across different imaging sites and equipment could help pinpoint specific attributes (e.g., intensity profiles, geometric distortions, texture patterns) that lead to significant inter-domain variations. Understanding these attributes would enable targeted improvements to mitigate their impact, such as specialized data augmentation techniques or domain adaptation strategies.

References

[1] Lijun An, Jianzhong Chen, Pansheng Chen, Chen Zhang, Tong He, Christopher Chen, Juan Helen Zhou, and B.T. Thomas Yeo. Goal-specific brain mri harmonization. *NeuroImage*, 263:119570, 2022. 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 2

[3] Mark A Brown and Richard C Semelka. *MRI: basic principles and applications*. John Wiley & Sons, 2011. 1

[4] Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018. 2, 4

[5] Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, and Aram Galstyan. Failure modes of domain generalization algorithms, 2021. 2

[6] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022. 2

[7] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020. 2

[8] Umang Gupta, Pradeep K. Lam, Greg Ver Steeg, and Paul M. Thompson. Improved brain age estimation with slice-based set networks, 2021. 1, 2, 4

[9] Umang Gupta, Pradeep K. Lam, Greg Ver Steeg, and Paul M. Thompson. Improved brain age estimation with slice-based set networks. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 840–844, 2021. 1, 2

[10] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2006. 2, 4

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021. 2

[12] Rafsanjany Kushol, Richard Frayne, Simon J. Graham, Alan H. Wilman, Sanjay Kalra, and Yee-Hong Yang. Domain adaptation ofnbsp;mri scanners asnbsp;annbsp;alternative tonbsp;mri harmonization. In *Domain Adaptation and Representation Transfer: 5th MICCAI Workshop, DART 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 12, 2023, Proceedings*, page 1–11, Berlin, Heidelberg, 2023. Springer-Verlag. 1

[13] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization, 2021. 2

[14] Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization, 2021. 2

[15] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation, 2019. 2

[16] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation, 2022. 2

[17] Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve, 2022. 2

[18] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016. 2

[19] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation, 2017. 2

[20] Yu-Wei Wang, Xiao Chen, and Chao-Gan Yan. Comprehensive evaluation of harmonization on functional brain imaging for multisite data-fusion. *NeuroImage*, 274:120089, 2023. 3

[21] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A survey, 2024. 2