Brain MRI Harmonization Literature Review

SeyedMostafa Ahmadi
ahmadism@yorku.ca

Medical Imaging Techniques
Professor Sadeghi-Naini
EECS5640

York University
Canada
April 2023

Contents

Consideration

In this literature review, we have tried to present a comprehensive of the trends in this literature. But before starting the literature review, there are some considerations to mention for the readers.

Structure

We first present a brief introduction to the topic in general and the causes and effects of the problem, and what the solution will bring to our lives in **Introduction** section (0.1). Then we discuss each paper in more detail. For each paper, we provide an **introduction and background** in the first section, then in **Methodology** and **Results** sections, we discuss the contributions and main results of each. For the **Methodology** section, we have also looked into the parent works as well to have a better understanding of the contributions. We more focus on understanding the ”what” that each paper tries to answer and the ”how” of their solution. Then we give our comments in the **Comments** section. At the end of exploring all the papers, we will provide a discussion on the future of this area in **Discussion** section (0.6).

Compactness

This project turned out to be a big project that took quite some time, and its original page count was four pages more than this, but we have tried to make it as compact as possible. As is the nature of journal papers, there are various interesting experiments in them, and mentioning all of them took a lot of space, so we decided to stick to the primary message of the papers and only explain them. We did not include any kind of table or figures to prevent them from taking up additional space, and we know this makes the report less interesting for the readers.

We made this document as brief as possible, which may have impacted the coherency, but we tried our best. There were some comments on the papers (in the Comments section) we removed because they were common in more than one paper, like the fact that all the first three paper only did their experiments on healthy patients, etc.

Comments for future students of the course

We regret that when we were choosing the papers to cover, we did not have the required knowledge to select the key papers of the topic, and although the presented papers are very interesting, some of them (the last two) are mostly following the previous works. This matter also affected us in running out of space to write since we had to cover more history of the papers.

Additionally, although choosing these four papers was amazing and taught us a lot of knowledge in this area, we do not recommend choosing more than three papers for future students. There is no upper limit on the chosen papers, but there is an upper limit on the page count of the document. Increasing the papers affects the presentation of the work, and it may seem inadequate, or you will lose a mark for being over the limit.

Disclaimer

We did not use any generative language model (ChatGPT or similar models) in this project, and everything is the result of human labor. But of course, we have used a grammatical helper tool (Grammarly) to help us reduce the mistakes.

0.1 Introduction

Magnetic Resonance Imaging is an important, non-invasive, and harmless imaging modality showing high contrast in soft tissues with a great resolution that has changed the game in diagnosis and treatment, especially its ability to produce detailed images in brain imaging for the diagnosis of neurological disorders, like strokes, tumors, and Alzheimer's disease. Due to brevity, we will not explain its benefits and applications and will dive into the problem definition.

There are plenty of researches claiming that differences in scanning sites create biases (non-biological differences) in MRI scans [1]. These differences can arise from a number of factors, some of them are differences in the scanner manufacturers, scanner coil, magnetic field strength, data acquisition protocols, and image reconstruction settings. There are various scanner manufacturers who create scanners with slightly different settings causing a slight difference in MRI scans of the exact same subject. Additionally, unalike coils and differences in the magnetic field strength can result in a difference in scans. Moreover, according to [2], the value of an MR image at a given voxel is determined by two dominant factors: the tissue properties and the scanner imaging protocol. The protocols refer to a set of instructions and settings, including parameters like FoV, TE, TR, etc. Also, the image reconstruction setting is a well-known problem that includes sampling data from the analog response.

To alleviate these effects, a new trend of research has begun to form under the name of harmonization. The main objective of harmonization methods is to distinguish between the biological variables of interest while making sure that the scanner (or site) used to acquire the data is not discernible. This is considered to be the most important factor in determining the effectiveness of the harmonization methods.

Harmonization has several benefits. There is a fabulous treasure of MRI datasets that can be explored for many things, such as demographical mega-analysis, life-span studies, etc., combining a number of these datasets. For example, Global Alzheimer's Association Interactive Network project has gathered MRI scans of more than half a million patients and is ready to use. Especially for learning-based studies, it is very important because it leads to empowering deep learning models by forcing the model to only focus on the biological variations rather than biases coming from non-biological sources. There has happened great efforts on MRI data analysis for a long time through academic and industrial research and even competitions (like this). Different tasks, like object detection, segmentation, classification, etc., have been explored through these efforts, where harmonization helps them better generalize by enabling models to use more diverse datasets.

A common goal for these efforts can be reaching the level of automated medical diagnosis or assistive diagnostic systems for physicians. This is important for the future of humankind, considering the increase in life expectancy and, as a result, the increasing prevalence of various mental illnesses like Alzheimer's disease or cancers, especially in elderly individuals. Besides, treatment in the very first stages of a disease is more beneficial for the economy, and sometimes treatments in the last stages are not effective and are just a waste of resources. Therefore, periodic screening for adults has high significance for individuals and governments for diagnostic purposes. And with limitations in the human resources of the medical community, moving towards assistive diagnostic systems for physicians seems obvious.

0.2 Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal [3]

The first paper we reviewed uses domain adaptation to attack the problem and is independent of the other methods in this review report that are either solely statistical or have statistical competitors, so we decided to explore it first. The paper takes advantage of domain adversarial neural networks [4], and through this adversarial scheme, learns its main tasks and, at the same time, unlearns the differences by the variations coming from the MRI scanning site (or scanner). The adversarial procedure used in this paper is an iterative procedure that has three steps. The main idea of the paper comes from [5], where they try to remove different biases from the datasets while classifying some factors. Thus, most of the analyses of this paper are following [5].

0.2.1 Methodology

They train networks consisting of three main parts (the first two have different names in their original paper): 1. Shared subnetwork 2. Task-specific subnetwork 3. Domain classifier

I named them this way since they have two tasks, and for each task, they have different networks with different structures. [1] The shared subnetwork gets the input image (3d image in age prediction and 2d image in segmentation) and gives a fully connected layer. This fully connected layer is connected to both the task-specific subnetwork and domain classifier. The task-specific subnetwork predicts the wanted results, i.e., either age or segmentation. And the domain classifier performs adversarial removal of the scanner effect.

In their training, they follow an iterative procedure with three steps on different parts of their network. First, they optimize the **shared subnetwork** and **task-specific subnetwork** for the main task. They try to minimize the mean square error (MSE) between the predicted age and the correct one in the age detection task and use a Dice loss for the segmentation task. This should be kept in mind that each batch only contains the data for one scanner. Second, they optimize the **domain classifier** to find the scanner based on the information in the fully connected layer. Based on different data coming from different scanners, the fully connected layer will contain information for the classification of the scanner, too (they prove this using a t-distributed Stochastic Neighbor Embedding [2] plot beautifully at their Fig. 6.). This domain classifier tries to find that data using a cross-entropy loss. In this step, the share subnetwork gets frozen, and they do not train that part. Third, they optimize the **shared subnetwork** so that the domain classifier gets confused and cannot identify the domain of the data. They typically use a log loss over all the probabilities of all the classes, where its best case is to have uniformly equal probabilities of the classes, and the goal is to confuse the domain classifier to meet the uniform distribution for its classifications.

The age prediction task takes the 3d volume of the brain MRI and tries to predict the age of the patient by learning a regression model. So, the task-specific subnetwork here models regression, and the shared subnetwork is a simple feature extraction that outputs the features in a vector for the use of the other two modules. The segmentation task, on the other hand, takes 2d slices of each 3d volume and segments them individually. The shared subnetwork is U-Net (Although Fig. 5 in their paper shows that the output of the shared subnetwork is the combination of the last layer and middle latent, in Table 6, we see that using only the last layer is the best. Honestly, I did not understand why they even suggested the use of the middle latent). The task-specific subnetwork here provides class probability maps for the segmentation, and they segment gray matter, white matter, and cerebrospinal fluid.

[1] One of the criticisms I have on the readability of their paper is this point. What I named ”shared subnetwork,” they call it ”feature extractor” in the section ”2.2. Network architecture” of their paper. However, in the section ”2.5. Segmentation task,” it is replaced with U-Net. Calling it a feature extractor does not seem to be correct at all.

[2] A non-linear and iterative dimensionality reduction method [6].

0.2.2 Results

They use three different datasets for this means, all of which are T1-weighted MRI scans of healthy patients. They separate training and test data in each dataset. For the age prediction network, they report better results in the main task compared to normal forms of training, where no unlearning has been performed on the model. They also add some other actions to be fairer in judgment, like trying to unbias the domains by selecting the overlapping ages, trying to confuse the domain classifier less often, etc. Also, the domain classifier performs close to random. For the segmentation network, they report a very small improvement in dice score for the main task, but they have removed the scanner effect successfully. However, comparing the models that train on a dataset and test on the same dataset, it is actually working worse.

0.2.3 Comments

Here we write seven comments that got into our minds as compactly as possible and one after the other (if any of them is not understandable, please contact me). The first problem with the learning-based models is that if we add a new dataset, we have to repeat the learning process from scratch. Furthermore, the biggest problem of this method is that the output (harmonized) is a latent vector, and it does not output a harmonized image familiar to the human eye of the operator or doctors. Moreover, they use a single control model for the main task on both datasets to compare with their approach (unlearning), which is unfair. Having a model for each dataset on the main task is the fairest scenario if they want to compare the improvement using unlearning. Another problem is that it needs labels for the main tasks that may not be available in large datasets. The additional neglected thing in their segmentation task is that the distribution of the patients should be similar in different domains (sites) as the error rises where the two distributions have the least overlap (But this is hard to find similar patients for segmentation). Also, they used 2d Slices for the segmentation task, which is problematic since the model can treat each slice differently, and the slices may end up not smoothed in the z-axis. But the good thing about the research is that they can remove other confounds like sex, race, .etc.

0.3 Harmonization of multi-site diffusion tensor imaging data [7]

The idea behind it comes from [8], where they propose a method to address the issue of batch effects in gene expression microarray data. Batch effects refer to non-biological variations in the data that happen because of differences in experimental conditions or technical factors, such as different reagents or operators, that affect the measurements in multiple samples. These batch effects affect the accuracy and reproducibility of data analysis significantly, and they try to remove these effects using a statistical model. [7] aims to remove the unwanted site effects in DTI [3] by mapping the batch effect to their problem and claim they have a lot in common.

0.3.1 Methodology

We discuss this paper more in detail compared to the other papers since we think it is important to understand it as the cornerstone of this literature comprehensively and touch on the mathematics behind it. We want to talk

[3] Diffusion tensor imaging (DTI) technique is a magnetic resonance imaging (MRI) technique used for the analysis of brain white matter microstructure and its integrity [9] (also see here). It is done using the characteristics of water diffusion in biological tissues. It is shown that water can not diffuse completely freely in biological tissues, and it is restricted by the cellular structure, for example, cell membranes, myelin sheaths, etc. Keeping this in mind, DTI aims to calculate a diffusion tensor for each voxel in the brain. Several parameters can be taken from the diffusion tensor, like fractional anisotropy and mean diffusivity. Fractional anisotropy (FA) is the measure of the degree of directionality of water diffusion in a certain voxel. For instance, if its value is 0, it means that water is diffusing in all directions, and if the value is 1, it means that water has completely anisotropic diffusion in the voxel (or diffuses in only one direction). Mean diffusivity (MD) shows the magnitude of water diffusion in a voxel and is just a scalar value.

about how their idea was formed and explain the course of these developments since it can be inspiring in any kind of research.

They have selected subsets of two datasets with data from two different sites, all of them related to healthy patients. For brevity, we will not mention the details of the datasets used and the work done to assure the quality of the selected subsets.The author, in their next paper [10], presents their competitor methods in the sequence of the ripening of the idea, probably detailing the evolution of the author's thinking to suggest the ComBat method. Here, we can somehow see this sequence as well.

The first mentioned model is Global Scaling which is a simple naïve method. In the method, they take the global mean among the whole dataset as a reference and find the mapping between the scanning center mean and global mean by fitting a simple linear model. In the model, they try to estimate the slope and y-intercept. This method can somehow remind us of histogram stretching mentioned in Question 7-b of the first experimental assignment of the course. The second method is functional normalization, where the authors use their previous work. Briefly speaking, they force all the histograms of $Y_{i,j}$ to be similar in distribution for each i and j ($Y_{i,j}$ is the scan$j$ from site i). The third method takes advantage of certain prior knowledge about the fractional anisotropy of cerebrospinal fluid. Knowing that FA values should always be close to zero (meaning that water is diffusing in all directions in cerebrospinal fluid), and if it is not like that, it is from the differences in scanning parameters and protocols, so the unwanted effects can be removed. They use a singular value decomposition to obtain the latent factors affecting the scan j in site i with unwanted variations and only use the strongest factor (w_1). And again, they use a linear regression model for all the white matter voxels. The fourth method is somehow the generalization of the previous model since they use more latent factors, but they obtain these latent factors differently. Just like the previous model, this one also models the effect of the latent factors linearly.

ComBat introduces a mixed regression model following the development of the idea and adds a multiplication error term for site effect to their model compared to the previous model. They assume measurement for a voxel follows the model in Equation 1 and site effect estimations are done using an empirical Bayes framework, then they suggest their harmonization removing the site effects.

$$y_{ijv} = \alpha_v + \mathbf{X}_{ij}\beta_v + \gamma_{iv} + \delta_{iv}\varepsilon_{ijv} \tag{1}$$

Where y_{ijv} is the measured voxel value for voxel v in scan j at site i, α_v is the average measurement of FA or MD for v, $\mathbf{X}$ is the matrix of interest covariates, β is the regression coefficients vector of $\mathbf{X}$, γ_{iv} is the unwanted effect of site i in voxel v, and $\delta_{iv}\varepsilon_{ijv}$ is the error term caused by site and scanner which has been expressed as the multiplication of pure site effect δ_{iv} (which is unwanted) and the regression error ε_{ijv}. We intentionally included the details of the formulation to use for referencing or criticizing in the next sections.

The method aims to estimate and remove the effect of γ_{iv} and δ_{iv} and, at the same time, to keep the intra-subject biological variability. For the first goal, they perform t-tests between two different sites (to determine if there is a significant difference based on their scans). And for the second goal, they check if different experiments can generate similar results (they use age as the result).

0.3.2 Results

First, they present the mean average difference for FA (or MD) measures after performing each harmonization method. As expected, RAVEL, SVA, and ComBat generate superior results compared to the two basic methods. They also measure the number of voxels significantly related to the site after performing each harmonization (using t-statistics with $p - value < 0.05$). Not surprisingly, most voxels are significantly related to the site before any harmonization. Again, RAVEL, SVA, and ComBat perform better in reducing the number of these voxels. Although, both of the previous results show that there is a small difference between SVA and ComBat.

In support of their second goal, they perform another experiment to see if the harmonization methods keep the biological differences at each site. They use the relationship between the age of the patient and the scan as a touchstone. They calculate the t-statistics between age and scan before and after harmonizations, and then they calculate the correlation of before and after. Each method that shows a better correlation before and after wins the game. Putting ComBat aside, the performance results are the exact opposite of the previous experiment. Global Scaling, Functional Normalization, and ComBat show superior results in comparison with RAVEL and SVA (SVA is the poorest). These results prove that ComBat preserves the biological variability and removes the effects of the site at the same time.

0.3.3 Comments

The paper is generally well-written and has great experiments supporting the idea, but there are some points that got into our minds. First, they only assume that the error term is additive and multiplicative. But it can be more complex, like something related to tissue type, $log()$ of the pixel value, etc. I have not seen any experiment on the type of model they explore. Second, they do not explore the reproducibility of the features using combat (this has been explored in the next paper we reviewed). Moreover, I do not have the expertise of this comment but have seen in some of the papers citing ComBat (like [1]) who have doubts about the similarity of the complexity of gene expressions and the features in MRI images. So, they think more experiments are required to check the performance of ComBat on all kinds of features.

0.4 Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics [11]

The paper we reviewed here uses traveling subjects between three MRI centers to evaluate the proposed methods and compare them with previous works, including ComBat. At first, we found this paper very interesting since they evaluated methods based on the scans of the same patients in different sites. The main message is that when we harmonize different patient scans from different sites, each patient has different features (indicators). Thus, comparison without considering these effects may be unfair. Using the traveling subject dataset, we can separate measurement bias from sampling bias which comes from patient differences.

0.4.1 Methodology

Their dataset consists of T1-weighted brain MRIs of 20 healthy patients scanned with three different procedures (however, one scan in the second one and three scans in the third procedure are missing). They also did a test-retest of 20 patients with procedure 3 of their dataset. The reason is that in the test-retest dataset, there is no measurement and sampling bias. Because using traveling subjects only removes the sampling bias caused by the difference in patients (like age, sex, .etc), and harmonization aims to remove measurement bias (caused by different scanners) as much as possible. And they do this to see how much they are successful in harmonization methods to get close to test-retest results.

They compare three methods of ComBat ([7]), TS-GLM([12]), and TS-ComBat (their proposed method). TS-GLM (they named it traveling-subject harmonization in [12]) is an extension of the generalized linear model (GLM) presented by [10] (in their paper, they call it the residuals method). GLM is a simple linear model for modeling site effects, and the parameters can be obtained using simple linear regression. TS-GLM separates the site bias (with site indicators) and patient-related bias (with patient indicators) and removes the site bias from the measurements. To give a bit more detail, a regression model is fitted on the data from each site to estimate the effects of sites on the data. The residuals of this model, which represent the measurement bias, are then used to harmonize the data across sites.

They have a quite minor change in TS-ComBat in comparison to ComBat, and that is the estimation of β_v in Equation 1. Site differences in ComBat can contain the sampling bias of the participants. But here in the TS-ComBat model, they only involve the same patient to prevent the sampling bias in the estimation of β_v. In this manner, the only present bias is the measurement bias which is purely related to the site.

0.4.2 Results

They have used some variables of FreeSurfer [4] for their output and the comparison of the results, like cortical thickness, cortical volume, and subcortical volume. For the comparisons, they have used Cohen's d[5] as their touchstone in their experiments. By this means, they calculate Cohen's d between the output of each pair of harmonized scans from procedures (1, 2, and 3) to check if it gets reduced compared to un-harmonized pairs. They perform a similar Cohen's d comparison on their test-retest pairs to check the reproducibility of biological features in each harmonization method.

They compared harmonization methods with the test-retest results to compare the measurement bias reduction. The results are expectable, and TS-GLM and TS-ComBat were superior to ComBat and are close to the test-retest Cohen's d results. But, looking at their reported charts, it looks like the TS-ComBat method is actually worse than TS-GLM! They also observed that Cohen's d is not zero for the test-retest experiment. They suspected this was due to other factors being present, like image analysis errors, individual errors, and measurement bias. But in the Comments section, we will discuss an important and neglected fact.

0.4.3 Comments

As it is obvious, the main contribution is from [12], and we believe there is a small contribution in this paper (adding that it is not that much better than TS-GLM). However, they have collected a really valuable dataset, and after all, scientific progress happens in small steps toward the correct answer! But there are some other considerations for this paper. The fact is that for the old collected datasets, there is not any demographical information available, and this way, we do not have any information about the sampling bias.

Another thing is that they are comparing the results in a cross-validation manner (Cohen's d) since there is no ground truth in these kinds of imaging. I will use the sampling topic discussed in the course material to support this comment. Any type of imaging is a digital sampling of analog signals (this analog signal is the ground truth). In this way, even two images from a patient taken by the exact same scanner with a small time difference are not the same and have their differences. This error can be prevented only by laying both 3d comb functions to fit exactly on each other, which is impossible. So, although this way of comparison seems approximately correct, it is not exactly. And this can be a reason why Cohen's d is not zero for the test-retest experiment.

[4] FreeSurfer is a software suite commonly used in the field of neuroimaging to analyze structural MRI brain scans. The output of FreeSurfer includes a number of different variables that represent various aspects of brain anatomy.

[5] Cohen's d is a statistical measure that represents the standardized difference between two means.

0.5 Goal-specific brain MRI harmonization [13]

The last paper we reviewed is quite new, and actually, no one has cited it yet! We used the number of citations as a measure of interest of the other researchers, but we don't have any idea about this paper. But after all, their concern for paying attention to the downstream application of the harmonization seemed pretty interesting, so we went for it. The downstream application in this work is Alzheimer's disease dementia prediction, and the fact that they have used unhealthy patients in their harmonization is interesting.

Their model in this paper is a minor extension to the conditional variational autoencoder (cVAE) presented in [14], having a number of fully connected layers added to cVAE's end. In [14], they used conditional variational autoencoders (a kind of generative adversarial network, like the one we described in Section 0.2) to create scanner-invariant feature representations for diffusion MRI. These features were used to reconstruct the images with minimal information on their original collection site. Their final results suggest that deep learning techniques can effectively create scanner-invariant representations for harmonizing MRI data since it is capable of finding non-linear relationships in an MRI scan of a particular site.

They look at the problem of harmonization this way: Trying to remove the **dataset** differences. This contains site differences, but it is not limited to them. For instance, they remove differences and biases caused by race as well. And at the end, they perform their harmonization technique and the competitors on ROI volumes of brains and compare the predictions based on different harmonization techniques.

0.5.1 Methodology

They used three datasets with T1-weighted brain MRI and matched one of them (Alzheimer's Disease Neuroimaging Initiative; ADNI) to the other two. For this matching, they found MRI scan pairs in each paired dataset ((i, j) pairs where scan i from ADNI matches with j scan from the other dataset) that were close together considering four factors of age, sex, mini mental state examination (MMSE) score, and clinical diagnosis (healthy, mild cognitive impairment, and Alzheimer's disease dementia) and formed a new matched dataset. They did this matching to find the closest pairs considering the mentioned factors to perform a fair experiment on each pair of datasets.

They use unmatched scans from each dataset to train the harmonization models, i.e., gcVAE (their proposed method), cVAE, and ComBat. After that, they harmonize the matched pairs and give them to a discriminator to see how much they are indistinguishable from each other after each harmonization technique. They train the discriminator (XGboost [15]) on the harmonized-unmatched MRI scans as well. And to see how much they do well in not removing the biological differences, they predict MMSE score and clinical diagnosis based on ROI volumes using a deep neural network on the result of harmonizations.

Their method is basically not new! Looking at Figure 4 in their paper, it might look like their model is a little bit different than cVAE (having an additional DNN layer), but it is not (The only difference is freezing and unfreezing the DNN network for the training phase). I know I am starting the criticism before the Comments section for this paper, but I have no choice since the explanation needs this clarification. They claim that gcVAE has an additional goal-specific DNN to the cVAE network. But they use the cVAE network similarly. They do the harmonization task separately, but for the prediction, they use the same DNN. Meaning that:

- For cVAE: When training for the harmonization, they freeze the DNN. For the MMSE and diagnosis prediction, they freeze the cVAE part.
- For gcVAE: When training for the harmonization, they do not freeze the DNN and train the predictor DNN as well.

0.5.2 Results

The results are quite close. Dataset prediction, after harmonization performed by the discriminator, predicts unharmonized data almost every time. But predicts cVAE and gcVAE with lower probability (a bit higher than random guess). For the performance of the clinical diagnosis, gcVAE outperforms other methods by an acceptable margin. But for the MMSE score, it actually is worse than cVAE and, in their first paired dataset, is worse than ComBat. In our opinion, it shows an unstable performance, and it is not much better than the other methods that are actually ”novel.”

0.5.3 Comments

Previously, we expressed our doubts about the novelty of the work. About the name of their paper (goal specific), the first paper we reviewed [3] does goal-specific harmonization as well, and this is not a novelty. But an interesting aspect of their research was pairing similar MRI scans. Although they did not discuss why they chose the chosen factors for pairing (and not some other factors). Also, it is not scalable for studying three or more datasets since the intersection of three or more sets gets smaller and smaller. Another good point of their research is that their output is harmonized image (and it is not a latent vector like [3]). But, since they do not freeze the DNN layer during the backpropagation, it might affect the harmonized image in a way that is not smooth or meaningful for the human eye. So, we were expecting them to include a sample for qualitative assessment of their work in the paper. But after all, it was interesting that they studied unhealthy patients as well as healthy patients. However, this was limited to the ROI regions and could not perform on the whole MRI scan.

0.6 Discussion

In the end, all of these studies have different strengths and weaknesses, as discussed. But there are some final comments we will make in this section. In the big picture and without a technical view, the best solution of this problem can be result of a cooperative work of the scanner manufacturers who can produce harmonization models for their scanners with different settings and compare to a reference that all the scanners in the world agree. But, due to the competitive nature of the business, this seems to be an unachievable mission, and the burden is on the researchers in academia.

Regarding the model for harmonizations, although the best fit may be a deep learning-based approach nowadays since it can predict non-linear relationships, we can move towards explainable models like the statistical models that are more sensible for humans and are not black boxes. But for explainable models, there might be different error distributions for each tissue type with totally different parameters, and this problem can be solvable by taking advantage of deep learning. For example, to separate the tissues for further analysis, we can use the recently introduced SAM [16] for accurate segmentation. This way, the models will be hybrid.

Also, to prevent training everything from scratch when the MRIs from a new scanner come to the database, we have to define a ”good” reference setting and try to harmonize every other scanning method to be something like that. But in the end, the world of learning lacks an important point that a good doctor can see a limited number of samples and learn the problem based on those, then generalize to the whole upcoming scans. But all of these harmonization methods need lots of data to harmonize and cannot perform in a good zero-shot learning ([17]) manner when there is a new scan from an unseen scanner.

0.7 Acknowledgement

In the end, I want to thank Professor Sadeghi, who permitted the literature review on this topic. Although my expertise is not in this particular field and my research focuses on computer vision for autonomous cars, I was very excited to learn new things, and I might end up diving into this field in the future path!

Bibliography

[1] S. A. Mali et al. “Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods”. In: *Journal of personalized medicine* 11.9 (2021), p. 842.

[2] J. L. Prince and J. M. Links. *Medical imaging signals and systems*. Vol. 37. Pearson Prentice Hall Upper Saddle River, 2006.

[3] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete. “Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal”. In: *NeuroImage* 228 (2021), p. 117689.

[4] E. Tzeng et al. “Simultaneous deep transfer across domains and tasks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4068–4076.

[5] M. Alvi, A. Zisserman, and C. Nellåker. “Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018, pp. 0–0.

[6] L. Van der Maaten and G. Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[7] J.-P. Fortin et al. “Harmonization of multi-site diffusion tensor imaging data”. In: *Neuroimage* 161 (2017), pp. 149–170.

[8] W. E. Johnson, C. Li, and A. Rabinovic. “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1 (2007), pp. 118–127.

[9] S. Mori and J. Zhang. “Principles of diffusion tensor imaging and its applications to basic neuroscience research”. In: *Neuron* 51.5 (2006), pp. 527–539.

[10] J.-P. Fortin et al. “Harmonization of cortical thickness measurements across scanners and sites”. In: *Neuroimage* 167 (2018), pp. 104–120.

[11] N. Maikusa et al. “Comparison of traveling-subject and ComBat harmonization methods for assessing structural brain characteristics”. In: *Human brain mapping* 42.16 (2021), pp. 5278–5287.

[12] A. Yamashita et al. “Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias”. In: *PLoS biology* 17.4 (2019), e3000042.

[13] L. An et al. “Goal-specific brain MRI harmonization”. In: *NeuroImage* 263 (2022), p. 119570.

[14] D. Moyer et al. “Scanner invariant representations for diffusion MRI harmonization”. In: *Magnetic resonance in medicine* 84.4 (2020), pp. 2174–2189.

[15] T. Chen and C. Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[16] A. Kirillov et al. “Segment anything”. In: *arXiv preprint arXiv:2304.02643* (2023).

[17] Y. Xian et al. “Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.9 (2018), pp. 2251–2265.